Notes from Survival Analysis

Cambridge Part III Mathematical Tripos 2012-2013

Lecturer: Peter Treasure

Vivak Patel

March 23, 2013

Contents

1 Introduction, Censoring and Hazard

1.1 Introduction

1. Aliases of Survival Analysis
   (a) In medicine: Survival Analysis
   (b) In engineering: Failure-time analysis
   (c) In general: Time-to-event Analysis
2. Framework
   (a) Scale: we need a scale to measure the duration of some event
   (b) Start Event: a clearly defined event when we start measuring with the scale
   (c) Event: A clearly defined event of interest

   Definition 1.1. *The time-to-event, T, is a random variable that measures the duration between the start event and the event. $T \geq 0$ and $T = 0$ at the start event.*
3. Properties
   (a) Generally, the scale is continuous, but is sometimes treated as discrete for convenience
   (b) Also, scale is usually time, but could be something that is appropriate (e.g. miles)
   (c) The scale flows left to right, so if an individual makes it to some t_2 which is greater than t_1, then the individual has made it past t_1

1.2 Censoring

Censoring occurs when we are unable to collect a complete set of data for a subject (i.e. we cannot record the time-to-event since the person drops out of the study, dies, the study ends etc.). It is impossible to avoid, so we must understand it and model it.

Definition 1.2. *1. Censoring occurs when we have incomplete information about the time-to-event.*

2. Time-to-censoring: C, is the duration between the start event and censoring

3. Right censoring: censoring which occurs before the event and after censoring the event has occurred but no information is collected

4. Uninformative Censoring: the act of censoring provides no information about the event time T

Remark 1.1. *If censoring is informative, then we must model it as a random event C, which complicates our analysis. It is sufficient to show that if censoring is stochastically independent of T, then C is uninformative.*

Example 1.1. *Suppose we are doing a clinical trial of a drug and our event occurs when the subject experiences symptom relief.*

1. *If censoring occurs because the trial is stopped according to a plan, then this is uninformative censoring for individuals who have not yet had an event.*
2. *If censoring occurs because an individual leaves the study owing to intolerable side effects, this can be informative censoring because the side-effects indicate that the drug is in effect and could relieve the individual's symptoms.*

1.3 Important Distributions

Definition 1.3. *Let T be a time-to-event random variable.*

1. *Survivor Function: $F(t) = \boldsymbol{P}[T > t]$*
2. *Distribution Function: $f(t) = -F'(t)$ or $F(t) = \int_t^\infty f(s)ds$*
3. *Hazard Function: the event rate at a time t conditioned upon observing time t without an event occurring: $h(t) = \frac{f(t)}{F(t)}$*
4. *Integrated Hazard Function:*
   (a) *$H(t) = \int_0^t h(s)ds$*
   (b) *$H(t) = -\log[F(t)]$ or $F(t) = \exp[-H(t)]$*

Example 1.2. *We consider the exponential family:*

1. *Suppose $F(t)$ is a non-negative, decreasing function on $[0, \infty)$ with $F(0) \leq 1$. The exponential survivor function is $F(t) = \exp[-\theta t]$ for $\theta > 0$*
2. *Suppose $h(t)$ is any non-negative function on $[0, \infty)$ such as $h(t) = \theta$. Then, by definition: $H(t) = \theta t$ and $F(t) = \exp[-\theta t]$*

Definition 1.4. *Families of Distributions:*

1. *Accelerated Life Family: if F is a survivor function and $\lambda > 0$, $F(t\lambda)$ defines an accelerated life family of distributions.*
2. *Proportional Hazard Family: if F is a survivor function and $k > 0$, then $F(t)^k = \exp(kH(t))$ defines a proportional hazards family of distributions.*

1.4 Notation

1. Let T_i be the time to event for an individual i
2. Let C_i be the time to censoring for an individual i
3. Let v_i be the visibility for an individual i. $v_i = \mathbf{1}[T_i \leq C_i]$. It is 1 if we see an event, 0 if the individual is censored.
4. Let $X_i = \min(C_i, T_i)$.

   Note 1.1. *There is a strict inequality for $v_i = 0$, i.e. $C_i < T_i$. If an individual is censored at time t we can either observe an event at time t or we do not ($T_i > C_i$)*

1.5 Likelihood Function in Survival Analysis

Suppose f, F, h, H all depend on some parameter θ:

1. If a person is censored ($v_i = 0$), then this subject contributes $F(x, \theta) = \mathbf{P}[T > x|\theta]$ to the likelihood function
2. If a person is not censored ($v_i = 1$), then this subject contributes $f(x, \theta) = \mathbf{P}[T = x|\theta]$

From this we have our likelihood and log-likelihood functions:

1. $L(\theta) = \prod_{i:v_i=1} f(x_i, \theta) \prod_{i:v_i=0} jF(x_i, \theta)$
2. $s(\theta) = \log[L(\theta)] = \sum_i v_i \log[f(x_i, \theta)] + (1 - v_i) \log[F(x_i), \theta]$

Lemma 1.1. $s(\theta) = \sum_i v_i \log[h(x_i, \theta)] - \sum_i H(x_i, \theta)$

Proof. By definition, $f(t) = h(t) \exp[-H(t)]$. Therefore,

$$\begin{aligned} s(\theta) &= \sum_i v_i \log[f(x_i, \theta)] + (1 - v_i) \log[F(x_i), \theta] \\ &= \sum_i v_i \log[h(x_i, \theta)] - v_i H(x_i, \theta) - (1 - v_i) H(x_i, \theta) \\ &= \sum_i v_i \log[h(x_i, \theta)] - \sum_i H(x_i, \theta) \end{aligned}$$

□

Example 1.3. *We compute the log-likelihood, MLE and Information for the exponential survivor function for which $h(t, \theta) = \theta$ and $H(t\theta) = t\theta$.*

Let d be the number of observed events and $X = \sum_i x_i$. Then:

1. *The log-likelihood: $s(\theta) = d \log[\theta] - X\theta$*
2. *The MLE: $\hat{\theta} = \frac{d}{X}$*
3. *The information at the MLE: $s''(\theta) = -\frac{d}{\hat{\theta}^2}$*

Notice that the information at the MLE for the exponential family depends only on how many individuals have experienced events.

2 Non- & Semi-parametric Inference & Testing

In parametric inference we assume some structure about an unknown parameter θ (e.g. we may know the dimension of the parameter space). In nonparametric inference, we assume θ is infinite dimensional or that the number of dimensions increases with each observed event.

Overview:

1. Kaplan-Meier Estimation, Proportional Hazards Modelling and Counting Processes (Nelson-Aalen Estimator) are methods for estimating the survivor or hazard functions
2. Log-rank test is a method for determining if two hazard functions are different
3. Model checking is about methods used to evaluate and improve our models

2.1 Kaplan-Meier Estimation

2.1.1 Introduction

1. Purpose: A non-parametric method for estimating the survivor function
2. Assumptions: we assume that events only occur at discrete times
3. Notation
   (a) We assume a fixed, sorted set of finite potential event times $\{a_0 < a_1 < \cdots < a_g\}$.
   (b) Let the number of individuals at risk (still without an event and not censored) at time a_j be r_j
   (c) The number of events at time r_j is $d_j \sim \text{Binomial}(r_j, q_j)$ where $q_j = \mathbf{P}[\text{event at } a_j | \text{ at risk at } a_j]$

2.1.2 Estimators

The Kaplan Meier Estimator is $\hat{F}(t) = \prod_{j:a_j \leq t}(1 - \hat{q_j}) = \prod_{j:a_j \leq t}(1 - \frac{d_j}{r_j})$

Derivation 2.1. *We derive the Kaplan-Meier Estimator*

1. *Note that there are no event gaps, i.e.* $\boldsymbol{P}[T \geq a_j] = \boldsymbol{P}[T > a_{j-1}]$
2. *Secondly,* $\boldsymbol{P}[T > a_j | T \geq a_j] = \boldsymbol{P}[\text{no event at } a_j | \text{at risk at } a_j] = 1 - q_j$
3. *Since q_j is conditional on r_j, which accounts for censored data, the censoring is uninformative*
4. *Suppose $a_{j-1} \leq t < a_j$, then*
$$F(t) = \boldsymbol{P}[T > t] = \boldsymbol{P}[T > a_0]\boldsymbol{P}[T > a_1 | T \geq a_1] \cdots \boldsymbol{P}[T > a_j | T \geq a_j]$$
5. *We estimate q_j with $\hat{q_j} = \frac{d_j}{r_j}$*

The variance of the Kaplan-Meier Estimator is $\mathbf{Cov}[\hat{F}(t)] = [F(t)]^2 \sum_{a_j \le t} \frac{1}{r_j} \frac{q_j}{1-q_j}$

Derivation 2.2. *We take advantage of the propagation of error formula to compute the variance, which approximately states that the error of a function $u(x)$ is equal to the first derivative of u squared times the squared error of x.*

1. *By the propagation of error formula, we have* $\boldsymbol{Cov}[u(X)] = [u'(\boldsymbol{E}[X])]^2 \boldsymbol{Cov}[X]$. *We apply this twice to* $\exp[\log[\hat{F}(t)]]$
2. *Note that* $\boldsymbol{Cov}[\hat{q_j}] = \frac{1}{r_j^2} \boldsymbol{Cov}[d_j] = \frac{1}{r_j^2}(r_j q_j (1-q_j)) = \frac{q_j(1-q_j)}{r_j}$
3. *Note that* $\boldsymbol{Cov}[\log[1-\hat{q_j}]] = \frac{1}{(1-q_j)^2} \frac{q_j(1-q_j)}{r_j} = \frac{1}{r_j} \frac{q_j}{1-q_j}$
4. *Therefore* $\boldsymbol{Cov}[\log[\hat{F}(t)]] = \boldsymbol{Cov}[\sum_{j:a_j \le t} \log(1-\hat{q_j})] = \sum_{j:a_j \le t} \frac{q_j}{r_j(1-q_j)}$
5. *Finally,* $\boldsymbol{Cov}[\exp[\log[\hat{F}(t)]]] = [\exp[\boldsymbol{E}[\log[\hat{F}(t)]]]]^2 \sum_{j:a_j \le t} \frac{q_j}{r_j(1-q_j)}$
6. *Letting* $\boldsymbol{E}[\log[\hat{F}(t)]] = \log[F(t)]$, *we have the stated variance.*

The estimator for the variance of the estimator is called Greenwood's formula (denoted s_0^2). It is obtained by approximating at time t with all events up to t:

$$s_0^2 = \hat{\mathbf{Cov}}[\hat{F}(t)] = [\hat{F}(t)]^2 \sum_{j:a_j \le t} \frac{1}{r_j} \frac{\hat{q_j}}{1-\hat{q_j}}$$

Assuming a standard normal distribution for $\hat{F}(t)$, the $1-\alpha$ confidence interval is:

$$\left[\hat{F}(t) - \Phi^{-1}\left(1-\frac{\alpha}{2}\right) s_0, \hat{F}(t) + \Phi^{-1}\left(1-\frac{\alpha}{2}\right) s_0\right]$$

2.1.3 Properties

1. Advantages of Kaplan-Meier Estimator
   (a) The estimator accounts for censoring, is maximum likelihood, and is nearly unbiased
   (b) KM Estimation can handle multiple events occurring at the same time (i.e. it can handle ties)
2. Disadvantages of Kaplan-Meier Estimator (mainly Greenwood's Formula)
   (a) The confidence interval is symmetric about $\hat{F}$ and may not be contained within $[0,1]$
   (b) By simulation, the confidence interval has a poor coverage probability
3. Fixes to Greenwood's Formula: instead we model $u(\hat{F}(t))$, estimate $\mathbf{Cov}[u(\hat{F}(t))]$, and compute:

$$\left[u^{-1}\{u\left(\hat{F}(t)\right) - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\}, u^{-1}\{u\left(\hat{F}(t)\right) + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\}\right]$$

Example 2.1. *Two possible functions for* $u(x)$ *are* $\log[x]$ *and* $\log[-\log[x]]$.

(a) For $u(x) = \log[x]$,

$$\boldsymbol{Cov}[u(\hat{F})] = \frac{1}{\boldsymbol{E}[\hat{F}(t)]}\boldsymbol{Cov}[\hat{F}(t)]$$
$$s_1^2 = \hat{F}(t)\sum_{j:a_j\leq t}\frac{1}{r_j}\frac{\hat{q_j}}{1-\hat{q_j}}$$

with confidence interval:

$$\left[\hat{F}(t)\exp\left(-\Phi^{-1}\left(1-\frac{\alpha}{2}\right)s_1\right), \hat{F}(t)\exp\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)s_1\right)\right]$$

(b) For $u(x) = \log[-\log[x]]$

$$s_2^2 = \frac{1}{\log[\hat{F}(t)]^2}\frac{1}{r_j}\frac{\hat{q_j}}{1-\hat{q_j}}$$

with confidence interval:

$$\left[\exp\left(\log\left[\hat{F}(t)\right]\exp\left(-z_{\alpha/2}s_2\right)\right), \exp\left(\log\left[\hat{F}(t)\right]\exp\left(z_{\alpha/2}s_2\right)\right)\right]$$

2.1.4 Profile Likelihood Curve for Confidence Intervals

1. We know that $\hat{F}$ is the maximum of the log-likelihood function s.
2. We can create a profile likelihood curve by solving for q_j when L is constrained by $F(t) = z \in [0, 1]$
3. From this curve, we can extract a $1-\alpha$ interval.

2.2 Proportional Hazards Modelling (Estimation)

2.2.1 Introduction

Note 2.1. *Proportional Hazards Modelling is sometimes called the Cox Model.*

1. Purpose: A semi-parametric method for estimating the hazard function.
2. Assumptions
   (a) We assume that the hazard function $h = \phi h_0$. We estimate ϕ parametrically, and we can estimate h_0 nonparametrically
   (b) We assume that there are no ties
3. Notation
   (a) The ith individual has explanatory variable $z^{(i)}$, and there is a parameter θ through which $z^{(i)}$ interacts with the hazard: $h_i(t) = h(t, z^{(i)}, \theta)$
   (b) By our assumptions, θ has a portion which we model parametrically β and a portion which we can model non-parametrically, ψ. Therefore, $h_i(t) = h(t, z^{(i)}, \beta, \psi) = \phi(z^{(i)}, \beta)h_0(t, \psi)$

i. ϕ is known as the hazard multiplier and is commonly assumed to be $\phi = \exp(\beta^T z)$
ii. h_0 is called the baseline hazard

(c) Suppose we have individuals $i = 1, \ldots, n$ each with $X_i = \min(T_i, C_i)$
i. Let events $j = 1, \ldots, d$ occur at time a_j.
ii. Since there are no ties, π_j, defined is the individual who has event at time a_j, is well specified
iii. Let $R_j = \{i | X_i \geq a_j\}$ be the risk set and $|R_j| = r_j$.

2.2.2 Partial Likelihood Functions

Definition 2.1. *A partial likelihood function is the likelihood function computed without all of the data (e.g. we leave out censored data).*

Remark 2.1. *There is significant theory which demonstrates that partial likelihoods behave like normal likelihoods. Therefore, we can construct confidence intervals, create hypotheses, apply Wilk's Lemma, etc.*

2.2.3 Estimator

We again use partial likelihood function (excluding censored data) and maximise it to determine β:

$$L(\beta) = \prod_j \frac{\phi(z^{\pi_j}, \beta)}{\sum_{i \in R_j} \phi(z^{\pi_i}, \beta)}$$

Derivation 2.3. *Recall that the likelihood function is*

$$L(\theta) = \prod_{i:v_i=1} f(x_i, \theta) \prod_{i:v_i=0} jF(x_i, \theta)$$

1. *The partial likelihood function, ignoring censored data, is then:*

$$L(\theta) = \prod_j f(x_{\pi_j}, \theta)$$

2. *Notice that by definition, $f(t) \propto h(t)$, and that*

$$\begin{aligned} f(x_{\pi_j}, \theta) &= \boldsymbol{P}[\pi_j \text{ has an event at } a_j | \pi_j \in R_j, \text{ there is an event at } a_j] \\ &\propto \phi(z^{\pi_j}, \beta) h_0(a_j, \psi) \\ &= \prod_j \frac{\phi(z^{\pi_j}, \beta)}{\sum_{i \in R_j} \phi(z^{\pi_i}, \beta)} \end{aligned}$$

2.2.4 Censoring

Suppose we have four individuals and the fourth individual is censored indicated by () after individual 3 has an event. Suppose the order of X_i is $3, (4), 1, 2$. Let the partial likelihood for this sequence of events be $L(3, 1, 2)$.

Individual 4 could have had an event in any of the following orders: $3, 4, 1, 2$, $3, 1, 4, 2$ or $3, 1, 2, 4$. The partial likelihood accounts for all of these events by $L(3, 1, 2) = L(3, 4, 1, 2) + L(3, 1, 4, 2) + L(3, 1, 2, 4)$. We say that the partial likelihood is **self consistent**.

2.2.5 Ties

Given a tie between two or more events we make the partial likelihood the sum of the partial likelihoods over all orders in which the tied individuals could have occurred if we had an infinitely exact time scale.

Example 2.2. *Suppose we have four individuals for whom we observe events. 3 has the first event. 4 and 1 tie. And 2 has the last event. Then:* $L(\beta) = L(3,1,4,2) + L(3,4,1,2)$.

2.3 Counting Processes & Nelson-Aalen Estimator

If we use proportional hazards, counting processes allow us to estimate H_0, the baseline hazard.

2.3.1 Counting Process

Definition 2.2. *Given individual* i *with* $X_i = \min(T_i, C_i)$ *and visibility* $v_i = \mathbf{1}[X_i = T_i]$.

1. *The counting process* $N_i(t)$ *is a random variable at each time* t *equal to the number of events for individual* i *up to and including time* t.

   Note 2.2. *In survival analysis,* $N_i(t) = \mathbf{1}[x_i \leq t, v_i = 1] \in \{0, 1\}$.

2. *Define* $dN_i(t) = N_i(t) - N_i(t-)$

3. *The history* H_t *of a process is everything we know about a process up to and including* t. H_{t-} *is the history just before* t.

4. *Intensity*

   (a) *The intensity* $\lambda(t)$ *is* $\lambda(t)\delta = \boldsymbol{P}[N(t+\delta) - N(t-)|H_{t-}]$

   (b) *The integrated intensity* $\Lambda(t) = \int_0^t \lambda(s)ds$ *so that* $\boldsymbol{P}[dN(t) = 1|H_{t-}] = d\Lambda(t)$

5. *A predictable process at time* t *is known given* H_{t-}

Lemma 2.1. *Let* N_i, *and* Λ *be defined as above. Then:*

1. $N_i(s) \leq N_i(t)$ *for* $s \leq t$

2. $\boldsymbol{E}[dN(t)|H_{t-}] = 1\boldsymbol{P}[dN(t) = 1|H_{t-}] + 0\boldsymbol{P}[dN(t) = 0|H_{t-}] = d\Lambda(t)$

3. $\Lambda(t)$ *is a predictable process.*

Note 2.3. *If we let* $M_i(t) = N_i(t) - \Lambda_i(t)$. *Then* $dM_i(t)$ *is a martingale with expectation 0. Therefore,* $N_i(t)$ *can be decomposed into a martingale and a predictable process.*

2.3.2 Counting Process in Survival Analysis

1. Intensity and Hazard: $\lambda(t)$ is $h(t)$ before an event happens, and is 0 after it happens. Therefore, we can write $d\Lambda = \mathbf{E}[dN_i(t)|H_{t-}] = Y_i(t)h_i(t)$.
   (a) $Y_i(t) = \mathbf{1}[T \geq t]$ is 1 before an event happens, and is predictable.
   (b) $\Lambda_i(t) = \int Y_i(s)h_i(s)ds = \int Y_i(s)dH_i(s)$
2. Notation
   (a) $N_+(t) = \sum_i N_i(t)$ is the data we have collected
   (b) $\Lambda_+(t) = \sum_i \int_0^t \int Y_i(s)dH_i(s)$ is what we want to estimate
   (c) $M_+(t) = N_+(t) - \Lambda_+(t)$ is a mean zero martingale

2.3.3 General Estimator

1. Using the fact that $M_+(t)$ is a mean-zero martingale, our general estimator will use $dN_+(t) = d\hat{\Lambda}_+(t) + 0$ to determine H_0.
2. Assumption: $h_i(t) = \phi(i)h_0(t)$ and $\phi(i)$ are known
3. The general estimator is:

$$\hat{H}_0(t) = \int_0^t \frac{dN_+(s)}{\sum_i Y_i(s)\phi(i)}$$

Derivation 2.4. *Using the fact that* $dN_+(t) = d\hat{\Lambda}_+(t) = \sum_i Y_i(t)\phi(i)d\hat{H}_0(t)$, *we have:*

$$d\hat{H}_0(t) = \frac{dN_+(t)}{\sum_i Y_i(t)\phi(i)}$$
$$\hat{H}_0(t) = \int_0^t \frac{dN_+(s)}{\sum_i Y_i(s)\phi(i)}$$

2.3.4 Nelson-Aalen Estmator

1. Assumptions: In addition to the assumptions for the general estimator, we assume that $h_j(t)$ are the same for all individuals, AND there are no ties.
2. Let $Y_+(t) = \sum_i Y_i(t)$. The Nelson-Aalen Estimator is then:

$$\hat{H}_0(t) = \int_0^t \frac{dN_+(s)}{Y_+(s)} = \sum_{j:a_j \leq t} \frac{1}{Y_+(a_j)}$$

Derivation 2.5. *By each assumption:*

(a) If h_i *are all the same then* $h_i(t) = h_0(t)$. *Therefore,* $\hat{H}_0(t) = \int_0^t \frac{dN_+(s)}{Y_+(s)}$

(b) Suppose events happen at times $a_1, \ldots, a_g$ *with no ties. Then* $dN_+(s) = \mathbf{1}[s = a_j]$ *for* $j = 1, \ldots, g$. *Moreover,* $Y_+(a_j) = r_j$ *the size of the risk set at time* a_j. *Therefore,* $\int_0^t \frac{dN_+(s)}{Y_+(s)} = \sum_{j:a_j \leq t} \frac{1}{Y_+(a_j)}$

3. Properties
   (a) It is easily generalised when individuals enter or leave the risk set
   (b) It is consistent with the Kaplan-Meier estimate for large n
   (c) Censoring is handled by Y_+
   (d) Calculations of the estimator of the variance of the estimator are straightforward
4. Dealing with Ties: Suppose we have a tie at a_j
   (a) One approach is to generalise the restrictions on $dN_+(t)$ and allow $\hat{H}_0 = \cdots + \frac{1}{r_{j-1}} + \frac{2}{r_j} + \frac{1}{r_{j+1}} + \cdots$
   (b) A second approach assumes time is continuous: $\hat{H}_0 = \cdots + \frac{1}{r_{j-1}} + \frac{1}{r_j} + \frac{1}{r_j - 1} + \frac{1}{r_{j+1}} + \cdots$

Remark 2.2. *$M(\infty) = y''$ is the martingale residual used in model checking.*

2.4 Log-Rank Test

2.4.1 Introduction

1. Purpose: Compares two survivor distributions with H_0 : both groups are identical
2. Notation
   (a) Groups are denoted by $i \in \{0, 1\}$
   (b) Survivor function at time $t = a_j$ in group i is denoted $F^{(i)}(a_j) = F_j^{(i)}$
   (c) Data: At time a_j, $r_j^{(i)}$ is the number at risk in group i and $d_j^{(i)}$ is the number of events in group i. Let $d_j = d_j^{(0)} + d_j^{(1)}$ and $r_j = r_j^{(0)} + r_j^{(1)}$.
3. Contingency Table at time a_j

	Group 0	Group 1	Total
Event	$d_j^{(0)}$	$d_j^{(1)}$	d_j
No Event	$r_j^{(0)} - d_j^{(0)}$	$r_j^{(1)} - d_j^{(1)}$	$r_j - d_j$
Total	$r_j^{(0)}$	$r_j^{(1)}$	r_j

4. Properties of Elements in the Contingency Table under H_0
   (a) $\mathbf{E}[d_j^{(0)}] = \frac{d_j}{r_j} r_j^0$
   (b) $\mathbf{Cov}[d_j^{(0)}] = \frac{r_j^{(0)} r_j^{(1)} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}$ (see Hypergeometric Distributions)

2.4.2 Test

1. Assumptions: We assume discrete event times $\{a_0 < a_1 < \cdots < a_g\}$ and the null hypothesis $F_j^0 = F_j^1$ for $j = 1, \ldots, g$.
2. Testing Parameters:
   (a) Statistic:: $z = \text{observed} - \text{expected} = \sum_{j=1}^{g} d_j^{(0)} - \sum_{j=1}^{g} \frac{d_j}{r_j} r_j^{(0)}$
   (b) Variance of Statistic: $s^2 = \sum_{j=1}^{g} \frac{r_j^{(0)} r_j^{(1)} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}$
   (c) Test Statistic: $\frac{z}{s} \sim N(0,1)$. We assume normality since we do not know its distribution.

   Derivation 2.6. *For the statistic:* $z_j = d_j^0 - \frac{d_j}{r_j} r_j^0$. *We simply sum over all values of* j.

   For the deviaiton:

$$\boldsymbol{Cov}[z] = \sum_j \boldsymbol{Cov}[z_j] = \sum_j \boldsymbol{Cov}[d_j^0] = \sum_j \frac{r_j^{(0)} r_j^{(1)} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}$$

3. Power:
   (a) The log-rank test works well if the survivor functions are in the same proportional hazards family
   (b) The log-rank test works poorly if the survivor functions have an intersection
4. Variants of the Log-Rank Test:
   (a) Some argue that the event times with larger risk sets should be weighted more
   (b) $z = \sum_{j=1}^{g} w_j z_j$, where $w_j = (r_j)^p$ usually with $p = 0, 0.5, 1.0$ usually.

2.4.3 Stratification

1. Purpose: suppose we are comparing two groups and want to control for another categorical variable
2. Statistic:
   (a) Create a contingency table for each level in the category, and compute z_j for each category
   (b) Compute z by summing over all z_j in all levels of categories.
3. Matched Pair Analysis: each stratification contains two individuals so that each stratum contributes at most 1 deviation value to the statistic.

2.4.4 Relative Risk

The relative risk is a measure of how to groups are. In the log-rank context, the relative risk is:

$$RR = \frac{\sum obs^1}{\sum exp^1} / \frac{\sum obs^0}{\sum exp^0}$$

2.5 Model Checking

2.5.1 General Model Checking and Survival Analysis

1. General model checking has the following process:
   (a) Generate and Fit the model
   (b) Analyse the residuals - most of the time these are naturally defined

   Remark 2.3. *In Survival Analysis, residuals do not have a natural definition and there is often more than one possible choice.*
2. Regression Model Checking
   (a) Residuals are typically observed value less the model value
   (b) Usually, the distribution of the residuals is exactly or approximately known

2.5.2 Cox-Snell Residuals

Proposition 2.1. *Suppose T is a continuous time-to-event with integrated hazard H. Let $U = H(T)$. Then $U \sim \exp(1)$*

Proof.

$$\begin{aligned}\mathbf{P}[U \leq t] &= \mathbf{P}[H(T) \leq t] = \mathbf{P}[T \leq H^{-1}(t)] \\ &= 1 - F(H^{-1}t) = 1 - \exp(-H(H^{-1}t)) \\ &= 1 - \exp(-t)\end{aligned}$$

□

1. Fit the model with $\hat{H}(t)$, and compute the Cox-Snell Residuals $y_i = \hat{H}(x_i)$
2. If $\hat{H}$ is close to H then y_i should be exponentially distributed by the proposition.
3. Compute the Kaplan-Meier curve for y_i and compare it to the expected exp(1) distribution
4. Determining the importance of a categorical variable:
   (a) Suppose the categorical variable has values A and B. Compute y_i^A using subjects in category A and y_i^B for subjects in category B.
   (b) Compute the KM curve for each group of residuals and compare them to the exp(1) distribution.

2.5.3 Cox-Snell Residuals and Censoring

Without censoring, $\mathbf{E}[y_i] = 1$, but when censoring occurs $\mathbf{E}[y_i] < 1$ because $x_i < T_i$, so it will pull the Kaplan-Meier curve for y_i down and to the left of the exp(1) curve. There are two ways we can fix this:

Note 2.4. $\boldsymbol{E}[v_i] = 1\boldsymbol{P}[X_i \geq T_i] \sim H(x_i)$.

1. Modified Cox-Snell: $y_i' = (1 - v_i) + \hat{H}(x_i)$. This accounts for censoring since with or without censoring, $\mathbf{E}[y_i'] = 1$.

2. Martingale Residuals: $y_i'' = v_i - \hat{H}(x_i)$. This is similar to the "observed-expected" residual, and $\mathbf{E}[y_i''] = 0$.

2.5.4 Martingale Residuals

1. Fit the model, excluding a (continuous) explanatory variable z, and calculate y_i''

2. Plot y_i'' against the values of z_i.

   (a) If the model, which excludes z, is sufficient, y_i'' should form an approximately horizontal line

   (b) If the model is incorrect, we should use nonparametric methods to compute $y'' = g(z)$ and include $g(z)$ in the model.

   Example 2.3. *If we use the cox model, and assume $\phi(\beta, \zeta) = \exp(\beta^T \zeta)$, we can test whether a variable z should be used in this model. If we see that it should, we have $\phi(\beta, \zeta, z) = \exp(\beta^T [\zeta \quad g(z)])$*

3. Compute y_i'' using the new model and ensure that the line is horizontal.

3 Advanced/Assorted Topics

3.1 Relative Survival Modelling (Estimation)

3.1.1 Introduction & Motivation

1. Motivation: Suppose we are observing two groups undergoing the same treatment, but have naturally different hazards. To account for this, we split $h(t) = h_B(t) + h_E(t)$.

   (a) h_B is the background hazard and refers to the hazard an individual has based on their circumstances, and it is usually known (governments usually publish this information)

   (b) h_E is the excess hazard, and is due to some factor that we are interested in learning about, such as a treatment

2. Distributions and Interpretations

   (a) We can compute the integrated background and excess hazards $H_B,\ H_E$ and the background and excess survivor functions $F_B,\ F_E$.

   (b) Background integrated hazard and survivor function have their natural interpretation

   (c) Excess integrate hazard and survivor function are fictitious functions we would expect to see if the background were not present

Example 3.1. *Suppose we treat a group of patients in Scotland and England, and see that survival is lower in Scotland than in England. By using relative survival modelling, we see that this is due to the fact that $h_B^{Scotland} > h_B^{England}$ while h_E is the same for both countries.*

3.1.2 Parametric Likelihood Based Estimation

1. Assumptions
   (a) We assume $h_B^i(t)$ and $H_B^i(t)$ are known
   (b) We assume $h_E^i(t)$ is the same for all subjections,and is $h_E(t)$
   (c) We assume we know the form (parametric) of $h_E(t, \theta)$.
2. The function s based on the Likelihood which we maximise to estimate θ and hence h_E is:
$$s(\theta) = \sum_i v_i \log[h_E(x_i, \theta) + h_B^i(x_i)] - \sum_i H_E(x_i, \theta)$$
3. Maximisation to determine θ requires a numerical approach.

   Derivation 3.1. *Using the second assumption, the partial likelihood functions is:*
$$\begin{aligned} s(\theta) &= \sum_i v_i \log[h_E(x_i, \theta) + h_B^i(x_i)] - \sum_i H_E(x_i\theta) + H_B^i(x_i) \\ &= \sum_i v_i \log[h_E(x_i, \theta) + h_B^i(x_i)] - \sum_i H_E(x_i\theta) \end{aligned}$$
   The second line follows from the fact that $H_B^i(x_i)$ are known constants and will not play a role in the maximisation.

Example 3.2. *Suppose we assume the form of $h_E(t) = \theta$ and so $H_E(t) = t\theta$. Then:*
$$s(\theta) = \sum_i v_i \log[\theta + h_B^i] - \theta \sum_i x_i$$
Let $X = \sum_i x_i$ and $d = \sum_i v_i$. Taking the derivative we have:
$$s'(\theta) = \sum_i \frac{v_i}{\theta + h_B^i} - X$$
With the added assumption that all $h_B^i = h$, we can set the derivative to 0 and solve to get $\hat{\theta} = \frac{d}{X} - h$. However, this is silly, but if we assume that $h_i \ll \theta$ we guess that we can have a similar solution:
$$\hat{\theta} = \frac{d}{X} - \epsilon \approx \frac{d}{X} - \frac{1}{d} \sum_i v_i h_B^i$$

Derivation 3.2. *We start by plugging in $s'(\hat{\theta})$ with $\hat{\theta} = \frac{d}{X} - \epsilon$*

1. *Plugging this in, rearranging, and setting it equal to 0, we have:*
$$\begin{aligned} s'(\hat{\theta}) &= \sum_i \frac{v_i}{h_B^i + \frac{d}{X} - \epsilon} - X \\ &= \frac{X}{d} \sum_i \frac{v_i}{1 + \frac{X}{d}(h_B^i - \epsilon)} - X \\ d &= \sum_i \frac{v_i}{1 + \frac{X}{d}(h_B^i) - \epsilon} \end{aligned}$$

2. *Expanding the function as a Taylor series, and solving for ϵ we have:*

$$d = \sum_i v_i(1 - \frac{X}{d}(h_B^i) - \epsilon)$$
$$d^2 = d\sum_i v_i - X\sum_i v_i h_B^i - X\epsilon\sum_i v_i$$
$$\epsilon = \frac{1}{d}\sum_i v_i h_B^i$$

3.1.3 Nonparametric Counting Process Estimation

1. Note that $H_E(t) = H(t) - H_B(t)$. By estimating $H(t)$ as we do in the counting process, we can estimate $H_E(t)$.

2. The estimator is:

$$\hat{H}_E(t) = \int_0^t \frac{dN_+(u)}{Y_+(u)} - \int_0^t \frac{\sum_i Y_i(u)h_B^i(u)du}{Y_+(u)}$$

   (a) The first term is the Nelson-Aalen Estimator

   (b) The second term is the weighted average of the background hazard by all individuals who have not had an event as of time t (since $Y_i = 1$ if $T > t$)

   Derivation 3.3. *Using the counting process:*

   (a) Recall: $dN_+(t) = d\Lambda_+(t) = \sum_i Y_i dH_E(t) + \sum_i Y_i h_B^i(t)$

   (b) Rearranging, we have: $dH_E(t) = \frac{dN_+(t) - \sum_i Y_I h_B^i(t}{Y_+(t)}$

   (c) We integrate to get the desired estimator

3. Beware that the first term is piecewise constant and the second is increasing, which may cause $H_E(t) < 0$ over some intervals.

   (a) We can fix this by: ignoring the relative parts, interpolating $\hat{H}_E$ between event times, or forcing $\hat{H}_E$ to be piecewise constant between event times.

   (b) However, these fixes do not work if H_E is negligible compared to H_B, or the disease preferentially affects a robust set.

3.2 Multiple Events Analysis: Marginal Modelling (Estimation)

3.2.1 Introduction

1. Multiple events analysis sutdies the case when more than one event is possible per individual; hence, a correlation exists between events.

2. Types of Multiple Events:

(a) Sequential: a person has events that occur in sequence (e.g. a person has headaches)

(b) Parallel: multiple time-to-events are being measured for the same individual (e.g. time until fillings in a subject's mouth fall out)

3. Dealing with Correlations

4. Frailty: assuming there is a hidden/latent variable (called Frailty) which creates a random effect between the correlated events. By conditioning on frailty, the events are conditionally independent.

5. Marginal Modelling: we analyse the data, ignoring dependence, and then we adjust for this dependence by altering the variance.

3.2.2 Jack-Knife

1. Purpose: The Jack-Knife is a method for computing the variance of an estimator when traditional methods fail.

2. Methodology

   (a) Let $J_i = \hat{beta}_i - \beta$ where $\hat{beta}_i$ is an estimate of β if we remove the ith observation. Thus, J_i is essentially a measure of the influence of the ith data point on the estimator.

   (b) $\hat{\mathbf{Cov}}[\hat{\beta}] = \frac{n-1}{n}\sum_i (J_i - \overline{J})(J_i - \overline{J})^T$ is the variance estimator

3. Properties

   (a) The Jack-knife is a robust, consistent and unbiased estimator when the data point left out is independent of the others

   (b) Jack-knife estimator for variance of normally distributed data is the usual variance estimator

3.2.3 Estimators

1. We assume a proportional hazards model and want to estimate β

2. The Newton-Raphson Method

   (a) Use the partial log-likelihood function to compute the score function $U(\beta)$ and information matrix $I(\beta)$.

   (b) Starting with some $\hat{\beta}^{(0)}$, we iteratively compute

   $$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + I^{-1}(\hat{\beta}^{(k)})U(\hat{\beta}^{(k)})$$

   (c) As $k \to \infty$, the iterated estimates converge to $\hat{\beta}$

3. Computing the Variance of the Estimator

   (a) We can compute the variance of the estimator using the Jack-knife method

(b) To compute $\hat{\beta}_i$, we can reiterate using the Newton-Raphson Method starting with $\hat{\beta}$ but by removing the ith data point altogether when computing the partial likelihood, score and information matrix.

4. In multiple event analysis, we leave out the entire subset of data associated with the ith subject instead of just the ith data point, since the subjects are independent.

3.2.4 Generic Data Structure

Example 3.3. *Suppose we are observing individuals who are normal. These individuals can either: progress into disease and then die, or die. Therefore there are three possible strata that exist:*

1. *Living → disease progression*
2. *Living → death*
3. *Disease Progression → death*

Also, suppose we have an explanatory variable indicating if a subject is treated (1) or is not (0). The following table demonstrates data which we might observe for two individuals:

Example Multi-Event Data Table

ID	Start	End	Status	Strata	Treatment
1	0	x_A^1	1	1	1
2	0	x_A^2	0	1	0
1	0	x_A^1	0	2	1
2	0	x_A^2	1	2	0
1	x_A^1	x_B^1	1	3	1

ID 1 The subject with ID 1 was observed (status = 1) to go through strata (1), and thus we could not observe him go directly to death, so this was censored information (status = 0 for strata = 2). In the last row, we do observe the individual die (status = 1) as the go from progression to death. And the individual is treated (treatment = 1)

ID 2 The subject with ID 2 was observed to go from living directly to death (status = 1 for strata = 2). Therefore, strata = 1 had to be censored (status = 0). The subject did not receive treatment (treatment = 0).

1. Row Headings:
   (a) A subject identifier which associates an event to a subject
   (b) Start time and End Time
   (c) Status, which corresponds to the visibility of the event
   (d) Stratification Identifier (the strata identifier indicates which event is observed – i.e. the specific progression from one state to another)
   (e) Explanatory variables
2. Each observation (event) has its own row

3.3 Frailty (Estimation)

3.3.1 Introduction

1. Motivation: often we are unable to include explanatory variables, especially if they are unknown to us. Typically, the effect of these variables are absorbed by the error term with randomisation. However, this does not always occur in survival analysis.

2. Problem Formulation:

   (a) Each subject has an unknown frailty (unknown explanatory variable), which we typically believe is non-negative

   (b) The higher an individual's frailty, the higher the risk of an event

3. Notation:

   (a) Let $U \geq 0$ be the frailty random variable

   (b) Distribution Functions for Individuals and Population

	Survivor Function	Hazard	Integrated Hazard
I:	$F_i(t) = F(t\|U = u_i)$	$h_i(t) = h(t\|U = u_i)$	$H_i(t) = H(t\|U = u_i)$
P:	$\bar{F}$	$\bar{h}$	$\bar{H}$

   (c) Since F is a straightforward probability, $\bar{F} = \mathbf{E}_U[F|U = u]$, but this is not true for hazards

   (d) Suppose $g(u)$ is some function. It's Laplace Transform is:

$$\widetilde{g}(\zeta) = \int g(u) \exp(-u\zeta) du$$

3.3.2 Proportional Frailty Model Estimator

Suppose we assume $h(t|U = u) = uh_0$. Then $\bar{F}(t) = \widetilde{g}(H_0(t))$.

Derivation 3.4. *By assumption, $H(t|U = 0) = uH_0(t)$. By definition, $F(t|U = u) = \exp[-uH_0(t)]$. Thus, if u has distribution $g(u)$*

$$\bar{F}(t) = \boldsymbol{E}_U[F(t|U = u)] = \int_0^\infty \exp[-uH_0(t)]g(u)du = \widetilde{g}(H_0(t))$$

Note 3.1. *We can often make $\boldsymbol{E}[U] = 1$ by absorbing any constant into the hazard, at time 0. However, as frail individuals are removed from the risk set at some later time t, $\boldsymbol{E}[U] < 1$. Therefore, individual hazards will decrease, on average, over time (the very frail ones die off early).*

Example 3.4. *Suppose $U \sim gamma(\psi, \psi)$, so that $\boldsymbol{E}[U] = 1$ and $\boldsymbol{Cov}[U] =$*

ψ^{-1}. *We have the following population distributions:*

$$\bar{F}(t) = \left(\frac{1}{1+\frac{H_0(t)}{\psi}}\right)^{\psi}$$
$$\bar{h}(t) = \frac{-\bar{F}'(t)}{\bar{F}(t)} = \frac{h_0(t)}{1+\frac{H_0(t)}{\psi}}$$
$$\frac{\bar{h}(t)}{h_0(t)} = \left(1+\frac{H_0(t)}{\psi}\right)^{-1}$$

1. *At $t = 0$, $H_0(t) = \int_0^0 h(s)ds = 0$. So we have that the third equation is 1, a consequence of $\boldsymbol{E}[U] = 1$.*
2. *As $t > 0$, the third equation will decreases since $H_0(t)$ is an increasing function. This correlates to $\boldsymbol{E}[U] < 1$ as time increases.*
3. *If ψ is small, $\boldsymbol{Cov}[U] = \psi^{-1}$ is going to be large. So the third equation will decrease rapidly as there will be individuals with very high frailties who will die very quickly.*

3.3.3 Influence of Unknown Variables in Proportional Hazards Model

1. The influence of Unknown Variables:
   (a) Suppose we have a simple treatment explanatory variable $z \in \{0, 1\}$ and we model the ith individual in group z using proportional hazards and proportional frailty: $h_i(t) = u_i^{(z)} e^{\beta z} h_0(t)$
   (b) Assume $U \sim gamma(\psi, \psi)$ and $u_i^{(z)}$ are i.i.d. Then:

$$\bar{h}^{(z)}(t) = \frac{e^{\beta z} h_0(t)}{1 + e^{\beta z} H_0(t)\psi^{-1}}$$
$$\frac{\bar{h}^{(1)}(t)}{\bar{h}^{(0)}(t)} = e^{\beta}\frac{1 + H_0(t)\psi^{-1}}{1 + e^{\beta} H_0(t)\psi^{-1}}$$

   (c) Notice that although individual hazard ratios are proportional over time, the population ratios change over time. So when we leave out an explanatory variable in proportional hazards, we do not capture the early behaviour of the population.
2. Coping with Unknown Variables
   (a) measure as many explanatory variables as possible and hope nothing is left out
   (b) Move away from proportional hazards completely, and use a model in which frailty will occur in the error (e.g. accelerated time family)

3.3.4 Inference of Frailty Variable

1. In general, $\bar{F}(t) = \widetilde{g}(H_0(t))$. The RHS of the equality can take on many forms of $\widetilde{g}$ and $H_0(t)$ and still achieve the same $\bar{F}(t)$ making inference difficult.

2. Moreover, we usually only have one observation (the event) for inference, making it nearly impossible to infer g

3. In multiple events, we can do better since we have multiple events per subject. However, with too few events per individual, convergence is difficult to achieve.

3.4 Cure (Estimation)

3.4.1 Introduction

1. Motivation
   (a) The cure model is a special case of frailty, in which some (unknown) fraction of the population is no longer at risk
   (b) Typically, a cure model's survivor function has levelled off, but this levelling off can have multiple explanations, such as:
      i. All subjects have very low risk
      ii. Some subjects have high risk, and some have a low risk
   (c) Therefore, choosing a cure model requires scientific input and evidence

2. Notation
   (a) The probability and individual is cured is $\pi_i = \pi(\beta, y_i)$, where y are explanatory variables and β are parameters, and π is a function taking values in $[0, 1]$.
   (b) The at risk/diseased fraction has distributions subscripted by a D: h_D, f_D, H_D, F_D. We do not observe these values
   (c) We do observe h_T, f_T, H_T, F_T for the whole population, where these parameters depend on parameters (γ) and explanatory variables (z_i).

3. Typical Parametrisation
   (a) $\pi_i = \left(1 + \exp[-\beta^T y_i]\right)^{-1}$ for an unknown β
   (b) $F_D^i = \exp\left[-\left(\exp(-\gamma^T z_i)t\right)^k\right]$ (Weibull) for unknown γ and k (k is an index in this context).
   (c) $H_D^i = \left(\exp(-\gamma^T z_i)t\right)^k$

3.4.2 Simple Model

1. The Simple Model
   (a) $F_T^i(t) = \pi_i + (1-\pi_i)F_D^i(t)$
   (b) $f_T^i(t) = (1-\pi_i)f_D^i(t)$ is an improper distribution
   (c) $s(\gamma, \beta, k) = \sum_i \log[(1-\pi_i)f_D^i(t)] + \sum_i (1-v_i)\log[\pi_i + (1-\pi_i)F_D^i(t)]$
2. Practical considerations when maximising s to determine γ, β, k
   (a) y and z may have common elements, so check $\mathbf{Cov}[y, z]$ to see what parameters may interfere.
   (b) We need to have many follow ups to ensure that an individual who is censored is actually cured.

3.4.3 Extensions to Simple Model

The first extension uses background mortality (hazard) to better model F_T. The second extension considers a semi-parametric proportional hazards modelling to better model F_T.

1. Adding Background Mortality
   (a) Let the vulnerable fraction have a background and excess hazard ($H_D(t) = H_B(t) + H_E(t)$).
   (b) Let the cured fraction have only a background hazard ($H_B(t)$)
   (c) Then, $F_T(t) = \pi F_B(t) + (1-\pi)F_B(t)F_E(t)$.
   (d) Compute f_T and the log-likelihood s accordingly and maximise to determine the unknown parameters.
2. Adding Proportional Hazards Modelling
   (a) Model: $h_D(t) = \exp[\gamma^T z]h_0(t)$
   (b) We then compute the distributions and log-likelihood:
      i. $F_D^i(t) = \exp\left(-e^{\gamma^T z}H_0(t)\right)$
      ii. $f_D^i(t) = \left(e^{\gamma^T z}h_0(t)\right)\exp\left(-e^{\gamma^T z}H_0(t)\right)$
      iii. $s = \sum_i v_i \log[(1-\pi_i)\left(e^{\gamma^T z_i}h_0(t)\right)\exp\left(-e^{\gamma^T z_i}H_0(t)\right)]$
      $+ \sum_i (1-v_i)\log[\pi_i - (1-\pi_i)\exp\left(-e^{\gamma^T z_i}H_0(t)\right)]$

      Note 3.2. *We can write* $\exp\left(-e^{\gamma^T z}H_0(t)\right) = [F_0(t)]^{\exp(\gamma^T z)}$
   (c) Because the baseline hazards do not cancel (since the hazard is only proportional to in h_D not h_T), h_0 and F_0 are nonparametric portions which cannot be directly maximised
   (d) In this situation, we use expectation maximisation to determine estimates for β and γ

3.4.4 Expectation Maximisation

Suppose we knew $q_i = \mathbf{1}$[individual i is vulnerable], then we could model the cured and diseased groups separately allowing us to determine β and γ. We would have the following log-likelihood for β and partial likelihood (based on proportional hazards models) for γ:

$$s_1(\beta) = \sum_i (1 - q_i) \log[\pi_i(\beta, y_i)] + q_i \log[\pi_i(\beta, y_i)]$$

$$L_2(\gamma) = \prod_{j:q_j=1} \frac{e^{\gamma^T z_j}}{\sum_{i \in R_j} e^{\gamma^T z_i}}$$

Note that we purposefully excluded the π_j notation for proportional hazards to avoid confusion, but this should be computed using the π_j notation. In expectation maximisation, we make an initial guess for $\widehat{q_i}$, compute the parameters, then update $\widehat{q_i}$ until we have convergence. Algorithm:

1. Guess $\widehat{q_i} = v_i$
2. Estimate β using the log-likelihood s_1
3. Estimate γ using the partial-likelihood L_2
4. Estimate F_0 and h_0 as well (Nelson-Aalen)
5. Updated $\widehat{q_i} = \mathbf{P}$[i is at risk|i is censored at x_i] $= \frac{\hat{F}(x_i)(1-\hat{\pi_i})}{\pi_i + \hat{F}(x_i)(1-\hat{\pi_i})}$
6. Repeat the steps using the new $\widehat{q_i}$ until $\widehat{q_i}$ converges

3.5 Empirical Likelihood

3.5.1 Introduction

1. Objective: We want to use nonparametric methods (empirical likelihood) to derive the survivor function and obtain point and interval estimates for time-to-event probabilities
2. Properties: the empirical likelihood function must satisfy the conditions of the survivor function $F(t)$:
   (a) Non-increasing
   (b) Non-negative
   (c) Bounded above by 1
3. Constructing the empirical likelihood from data:
   (a) Individual contributions to the likelihood:
      i. If $v_i = 1$, then $\mathbf{P}[T = x_i] = F(x_i) - F(x_i-)$
      ii. If $v_i = 0$, and $T > x_i$ (right censored), then $\mathbf{P}[T > x_i] = F(x_i)$
      iii. If $v_i = 0$, and $T \leq x_i$ (left censored), then $\mathbf{P}[T \leq x_i] = 1 - F(x_i)$

iv. If $v_i = 0$, and $T \in [x_i^L, x_i^U]$ (interval censored), then $\mathbf{P}[T \in [x_i^L, x_i^U]] = F(x_i^L) - F(x_i^U)$

(b) Common simplifications

i. We assume we only have events or right censoring

ii. If $v_i = 1$ then $\widehat{F}(x_i-) =$ $\widehat{F}$(the largest preceding time when an event occured)

iii. If $v_i = 0$ then $\widehat{F}(x_i) =$ $\widehat{F}$(the largest preceding time when an event occured)

3.5.2 Derivation of Kaplan-Meier

1. Assuming the simplifications, and:

   (a) Suppose $a_1 < a_2 < \cdots < a_g$ are event times, and let $a_0 = 0$, $a_{g+1} = \infty$

   (b) Suppose $d_j \geq 1$ events occur at the event times with subscript $j = 1, \ldots, g$.

   (c) Suppose c_j individuals are censored between $[a_j, a_{j+1}]$

2. The likelihood functions are then:

   (a) The likelihood function:

$$\begin{aligned} L &= \prod_{j=1}^{g} [\mathbf{P}[T = a_j]]^{d_j} [\mathbf{P}[T > a_j]]^{c_j} \\ &= \prod_{j=1}^{g} [F(a_{j-1}) - F(a_j)]^{d_j} [F(a_j)]^{c_j} \end{aligned}$$

   (b) The log-likelihood function:

$$s = \sum_{j=1}^{g} d_j \log[F(a_{j-1}) - F(a_j)] + c_j \log[F(a_j)]$$

   (c) The partial derivative of the log-likelihood:

$$\frac{\partial s}{\partial F(a_j)} = \frac{d_j}{F(a_j) - F(a_{j+1})} + \frac{c_j}{F(a_j)} - \frac{d_j}{F(a_{j-1}) - F(a_j)}$$

3. We have the Kaplan-Meier by the following procedure:

   (a) Set the partial derivative to 0 for event time a_g and solve to get: $\hat{F}(a_g) = \frac{c_g}{c_g + d_g} \hat{F}(a_{g-1})$

   (b) Let $r_g = c_g + d_g$ and $r_j = r_{j+1} + c_j + d_j$ giving: $\hat{F}(a_g) = (1 - \frac{d_g}{r_g}) \hat{F}(a_{g-1})$

   (c) Prove by induction: $\hat{F}(0) = 1$, then

$$\hat{F}(a_j) = \left(1 - \frac{d_g}{r_g}\right) \hat{F}(a_{j-1})$$

3.5.3 Constrained Maximisation for Interval Estimation

Note 3.3. *We can generate an interval for each F(t) by creating a profile curve* $F(t) = z \in [0, 1]$. *This is the same as before, but here we describe the method of doing it.*

Method of Lagrangian Multipliers.

1. Suppose we have the constraint that $F_k = z$
2. Let the Lagrangian $S(\lambda) = \sum_{j=1}^{g} d_g \log[F_{j-1} - F_j] + c_j \log[F_j] + \lambda(\log[F_k] - \log[z])$
3. We want to maximise S with respect to all λ which also satisfies the constraint.
4. Using the recurrence relationship used to derive the KM Estimator, we note:
   (a) Starting with $\widetilde{F_k} = z$, $\widetilde{F_j} = (1 - \frac{d_j}{r_j})\widetilde{F_{j-1}}$ for $j > k$
   (b) Starting with $\widetilde{F_0} = 1$, we need to find the λ which satisfies $\widetilde{F_k} = 0$ using the recurrence $\widetilde{F_j} = (1 - \frac{d_j}{r_j} + \lambda)\widetilde{F_{j-1}}$ for $j \leq k$

Remark 3.1. *In practice, it is easier to simply start with values of* λ *and compute the* z *to which they correspond.*

3.6 Schoenfeld Residuals (Model Checking)

3.6.1 Introduction

1. In the Cox Model, we want to estimate β where $h^i(t) = h_0(t) \exp[\beta^T z]$. Cox-Snell and Martingale residuals allowed us to evaluate the plausibility of $\beta^T z$ relationship and replace it with $\beta^T g(z)$ if necessary. Schoenfeld residuals allow us to determine the time dependence of β.
2. Notation
   (a) Let $Y_i(t) = \mathbf{1}[i \in R_t]$
   (b) Let $w_i(t) = Y_i(t) \exp[\beta^T z_i]$
   (c) Let $\bar{z}_j(\beta) = \frac{\sum_i w_i z_i}{\sum_i w_i}$ be the weighted average of the explanatory variables in the risk set at time a_j
3. Assumptions
   (a) We assume the Cox (Proportional Hazards) Model
   (b) We assume there are no ties and d events occurring at time $a_1 < a_2 < \cdots < a_d$

4. Recall from the Cox Model that the likelihood, log-likelihood, score and information are:

$$\begin{aligned} L(\beta) &= \prod_j \frac{\exp[\beta^T z^{\pi_j}]}{\sum_i w_i(a_j)} \\ s(\beta) &= \sum_j \beta^T z^{\pi_j} - \log[\sum_i w_i(a_j)] \\ U(\beta) &= \sum_j z^{\pi_j} - \frac{\sum_i \partial_\beta w_i(a_j)}{\sum_i w_i(a_j)} \\ &= \sum_j z^{\pi_j} - \bar{z}_j(\beta) \\ I(\beta) &= -\partial_\beta U(\beta) = \sum_j \partial_\beta \bar{z}_j(\beta) \end{aligned}$$

3.6.2 Schoenfeld Residual and Applications

The Schoenfeld residual is defined based on the score function, by noting that if β does not have a time dependence $U(\hat{\beta}) = 0$.

Definition 3.1. *Let the Schoenfeld Residual be* $s_j(\beta) = z^{\pi_j} - \bar{z}_j(\beta)$. *Let* $s_j^*(\beta) = I^{-1}(\beta) s_j(\beta)$

Lemma 3.1. *Therefore,* $U(\beta) = \sum_j s_j(\beta)$ *and* $\boldsymbol{E}[s_j^*(\beta)] = \theta g(a_j)$

Proof. The first equality follows from the definitions of U and s_j. To prove the second equality (with hand-waving):

1. Let $\beta(t) = \beta_0 + \theta g(t)$. Notice that adding or subtracting a constant to $g(t)$ will get absorbed in proportional hazards modelling by the exponential term. So we can scale the function to ensure that $\mathbf{E}[\hat{\beta}] = \beta_0$
2. Given that the correct form of the parameter is $\beta(t)$, we have that:
   (a) $\mathbf{E}[s_j(\beta_0)] = \bar{z}_j(\beta(a_j)) - \bar{z}_j(\beta_0) = \bar{z}_j(\beta_0 + \theta g(a_j)) - \bar{z}_j(\beta_0)$
   (b) By Taylor Expansion:

$$\mathbf{E}[s_j(\beta_0)] = \bar{z}_j(\beta_0 + \theta g(a_j)) - \bar{z}_j(\beta_0) = \sum_j \partial_\beta \bar{z}_j|_{\beta_0} \theta g(a_j) = I(\beta_0)\theta g(a_j)$$

   (c) Therefore, $\mathbf{E}[s_j^*(\beta_0)] = \mathbf{E}[I^{-1} s_j(\beta_0)(\beta_0)] = \theta g(a_j)$
3. Using $\hat{\beta}$ as an estimator for β_0, we have the desired property.

□

Application: We can compute $s_j^*(\hat{\beta})$ and plot it against a_j.

1. If β does not have a time-dependence, then the line will be horizontal
2. IF β hat does have a time-dependence, we can compute $\theta g(t)$ using non-parametric methods.

3.7 Planning Experiments: Determining size of a Study

3.7.1 Introduction

1. Objective: To determine how many subjects we need in a study to conclude that one treatment is better than another, given a significant level α and power $1-\beta$

2. Assumptions

   (A0) The test statistic, U, is normally distributed under the null and alternative hypotheses

   (A1) We assume the hazards under the two treatments (0,1) are proportional

   (A2) There are no ties in the data

   (A3) Risk sets between both treatments remain equal at all times

   (A4) We assume survivor functions under both treatments are exponential

3. Revision of Testing Treatments

   (a) Hypothesis: we are hoping and expecting $\mu_1 > \mu_2$, else we would not do this test. By (A0):

   H_0: $U \sim N(\mu_0, \sigma_0^2)$
   H_1: $U \sim N(\mu_1, \sigma_1^2)$

   (b) Test Characteristics

      i. $\mathbf{P}[\text{rejecting } H_0 | H_0 \text{ is true}] = \alpha$ called the significance
      ii. $\mathbf{P}[\text{rejecting } H_0 | H_1 \text{ is true}] = 1-\beta$ called the power

4. Under the Log Rank Test with $U = \frac{z}{s}$, by assumption (A0)

   H_0: $U \sim N(0,1)$
   H_1: $U \sim N(\mu_1, \sigma_1^2)$, but the parameters are unknown

5. Notation: Let $d = \sum_j 1$

3.7.2 Sample Size Inequality

We first note that to achieve a significance level of α under H_0, we have for some critical value C:

$$\begin{aligned}
\alpha/2 &= \mathbf{P}[U > C | H_0] \\
&= \mathbf{P}[\frac{U-\mu_0}{\sigma_0} > \frac{C-\mu_0}{\sigma_0} | H_0] \\
&= 1 - \Phi\left(\frac{C-\mu_0}{\sigma_0}\right) \\
C &= \sigma_0 \Phi^{-1}(1-\alpha/2) + \mu_0
\end{aligned}$$

Noting that we expect $\mu_1 > \mu_2$, to achieve the power level $1-\beta$ under H_1, we have for the same critical value C:

$$\begin{aligned} 1-\beta &\leq \mathbf{P}[U > C|H_1] \\ &\leq \mathbf{P}[\frac{U-\mu_1}{\sigma_1} > \frac{C-\mu_1}{\sigma_1}|H_1] \\ &\leq 1-\Phi\left(\frac{C-\mu_1}{\sigma_1}\right) \\ C &\leq \sigma_1\Phi^{-1}(\beta)+\mu_1 \end{aligned}$$

Combining the two results, we have the sample size inequality:

$$\begin{aligned} \mu_1-\mu_0 &\geq \sigma_0\Phi^{-1}(1-\alpha/2)-\sigma_1\Phi^{-1}(\beta) \text{ or} \\ \mu_1-\mu_0 &\geq \sigma_0\Phi^{-1}(1-\alpha/2)+\sigma_1\Phi^{-1}(1-\beta) \end{aligned}$$

3.7.3 Parameters under H_0

We want to compute the form, expectation and variance of U under H_0. Under assumptions (A2) and (A3) we that $d_j = 1$ and $r_j^0 = r_j^1$. Under H_0, we have that:

1. $\mathbf{E}[d_j^0] = \mathbf{P}[d_j^0 = 1|H_0] = 0.5$. Under H_0 there is no difference between the treatments, so either an event occurs under treatment 1 or treatment 0, and both should have the same probability.
2. $\mathbf{Cov}[d_j^0] = \mathbf{E}[(d_j^0)^2] - (\mathbf{E}[d_j^0])^2 = 1^2\mathbf{P}[d_j^0 = 1|H_0] - 0.25 = 0.5 - 0.25 = 0.25$
3. $U = \frac{(\sum_j d_j^0) - d/2}{\sqrt{d/4}}$
4. $\mathbf{E}[U] = 0 = \mu_1$ as desired
5. $\mathbf{Cov}[U] = 1 = \sigma_0^2$ as desired

Derivation 3.5. *Recall from the Log-Rank test that the form of* $U = \frac{\sum_j obs-exp}{\sqrt{\boldsymbol{Cov}[\sum_j obs-exp]}}$. *Therefore:*

1. $U = \frac{\sum_j d_j^0 - 1/2}{\sqrt{\sum_j \boldsymbol{Cov}[d_j^0]}}$ *which is what we are looking for*
2. $\boldsymbol{E}[U] = \frac{2}{\sqrt{d}}(\sum_j(\boldsymbol{E}[d_j^0]) - d/2) = 0$
3. $\boldsymbol{Cov}[U] = \frac{4}{d}\sum_j \boldsymbol{Cov}[d_j^0] = 1$

3.7.4 Parameters under H_1

We want to compute the form, expectation and variance of U under H_1. Under assumption (A1), $h_0(t)/h_1(t) = \lambda$. Under H_0, $\lambda = 1$, and under H_1, $\lambda > 1$. Under H_1 we have that:

1. $\mathbf{P}[d_j^0 = 1|H_1] = \frac{\lambda}{1+\lambda}$
2. $\mathbf{E}[d_j^0] = \frac{\lambda}{1+\lambda}$ and $\mathbf{Cov}[d_j^0] = \frac{\lambda}{(1+\lambda)^2}$

3. $\mathbf{E}[U] = \sqrt{d}\frac{\lambda-1}{\lambda+1}$ and $\mathbf{Cov}[U] = \frac{4\lambda}{(1+\lambda)^2}$

Derivation 3.6. *Under H_1 and using the form of U from the previous section:*

1. *Recall that $\boldsymbol{P}[\text{event in group } i \text{ at time } j|H_k] \propto r_j^i h_i(a_j)$. Therefore:*

$$\begin{aligned}
\boldsymbol{P}[grp^1|H_1] &\propto r_j^1 h_1(a_j) \\
\boldsymbol{P}[grp^0|H_1] &\propto r_j^0 h_0(a_j) = r_j^1 h_1(a_j)\lambda \\
\boldsymbol{P}[grp^0|H_1] &= \frac{r_j^1 h_1(a_j)\lambda}{r_j^1 h_1(a_j)\lambda + r_j^1 h_1(a_j)} = \frac{\lambda}{1+\lambda}
\end{aligned}$$

2. *For the expectation and variance of d_j^0 we have:*

$$\begin{aligned}
\boldsymbol{E}[d_j^0] &= 1\boldsymbol{P}[grp^0|H_1] = \frac{\lambda}{1+\lambda} \\
\boldsymbol{Cov}[d_j^0] &= \boldsymbol{E}[(d_j^0)^2] - (\boldsymbol{E}[d_j^0])^2 \\
&= \frac{\lambda}{1+\lambda} - \frac{\lambda^2}{(1+\lambda)^2} \\
&= \frac{\lambda}{(1+\lambda)^2}
\end{aligned}$$

3. *For the expectation and variance of U we have:*

$$\begin{aligned}
\boldsymbol{E}[U] &= \frac{2}{\sqrt{d}}(\sum_j (\boldsymbol{E}[d_j^0]) - d/2) \\
&= 2\sqrt{d}\left(\frac{2\lambda}{2+2\lambda} - \frac{1+\lambda}{2+2\lambda}\right) \\
&= \sqrt{d}\frac{\lambda-1}{\lambda+1}
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{Cov}[U] &= \frac{4}{d}\sum_j \boldsymbol{Cov}[d_j^0] \\
&= \frac{4}{d}\sum_j \frac{\lambda}{(1+\lambda)^2} \\
&= \frac{4\lambda}{(1+\lambda)^2}
\end{aligned}$$

3.7.5 Sample Size

Using the sample size inequality, and assumptions (A0), (A1), (A2), & (A3), we have that:

$$d \geq \left(\frac{\lambda+1}{\lambda-1}\right)^2 \left[\Phi^{-1}(1-\alpha/2) + \frac{2\sqrt{\lambda}}{\lambda+1}\Phi^{-1}(1-\beta)\right]^2$$

Derivation 3.7. *Plugging into the sample size inequality with* $\mu_1 = \boldsymbol{E}[U|H_1]$ *and* $\sigma_1 = \sqrt{\boldsymbol{Cov}[U|H_1]}$

$$\sqrt{d}\frac{\lambda-1}{\lambda+1} - 0 \geq \Phi^{-}(1-\alpha/2) + \frac{2\sqrt{\lambda}}{1+\lambda}\Phi^{-}(1-\beta)$$

$$d \geq \left(\frac{\lambda+1}{\lambda-1}\right)^2 \left[\Phi^{-1}(1-\alpha/2) + \frac{2\sqrt{\lambda}}{\lambda+1}\Phi^{-1}(1-\beta)\right]^2$$

To compute the sample size n, we suppose at time a_m the survival probability in group 0 is π_0 (known) and in group 1 is π_1 (unknown). We note that by assumption (A4), $\pi_1 = \pi_0^{1/\lambda}$.

Derivation 3.8. *Assumption (A4) says that the survivor functions are exponentially distributed. So the density is exponential. Adding assumption (A1), we have for the densities:*

$$\log[\pi_0] = -a_m \lambda H_1$$
$$\log[\pi_1] = -a_m H_1$$

Therefore, $\pi_0 = \pi_1^{\lambda}$.

The sample size given this set up is then:

$$n = \frac{d}{2 - \pi_0 - \pi_1}$$

Derivation 3.9. *Note that* $d_m^i = n(1-\pi_i)$ *where* n *is the number of subjects in total. Then:*

$$d = d_m^0 + d_m^1 = n(2 - \pi_0 - \pi_1)$$

The result follows by rearrangement.

Hypothesis Tests and Confidence Regions Using the Likelihood

F. P. Treasure

13-May-05
3020a

Maximum Likelihood:

The likelihood function for a model parameterised by θ (a vector in a p-dimensional space Θ) given observed data vector x is $L(\theta|x)$ which we abbreviate to $L(\theta)$. The log-likelihood $S(\theta)$ is defined by $S(\theta) := \log L(\theta)$.

The value of θ which maximizes $S(\theta)$ is $\hat{\theta}$ and is the *maximum likelihood estimate* of θ :

$$S(\hat{\theta}) = \max_{\theta \in \Theta} S(\theta) \ .$$

$\hat{\theta}$ is normally found by solving the score equations $S'(\hat{\theta}) = 0$, where a prime denotes differentiation.

If θ is constrained to a q-dimensional subspace Θ_0 of Θ then the value of θ maximizing $S(\theta)$ in that subspace is $\tilde{\theta}$:

$$S(\tilde{\theta}) = \max_{\theta \in \Theta_0} S(\theta)$$

and $\tilde{\theta}$ can generally be found using Lagrange multipliers.

Hypothesis tests

The (non-negative) difference $S(\hat{\theta}) - S(\tilde{\theta})$ is the reduction in the log-likelihood due to constraining the space from Θ to Θ_0 . If the data does not support $\theta \in \Theta_0$ then we would expect $S(\hat{\theta}) - S(\tilde{\theta})$ to be relatively large and vice-versa.

Wilks's lemma tells us that:

$$2\left[S(\hat{\theta}) - S(\tilde{\theta})\right] \sim \text{chisquare}\,(p-q)$$

(provided $\theta \in \Theta_0$) and so we can compare $2\left[S(\hat{\theta}) - S(\tilde{\theta})\right]$ with the chisquare $(p-q)$ distribution and obtain a size α hypothesis test for the null hypothesis $\theta \in \Theta_0$:

$$\text{accept hypothesis } \theta \in \Theta_o \text{ if } S(\hat{\theta}) - S(\tilde{\theta}) \leqslant \frac{1}{2} C_{p-q,1-\alpha} \qquad (1)$$

where $C_{m,\gamma}$ is the γth quantile of a chisquare(m) distribution.

Confidence Regions and Intervals

A p-dimensional $1-\alpha$ confidence *region* can be constructed by letting Θ_0 consist of the single point θ_0 and including θ_0 in the confidence region if the hypothesis test $\theta \in \Theta_0$ – equivalently: $\theta = \theta_0$ – is not rejected by rule (1). The confidence region is therefore (noting here that $q = 0$ and $\tilde{\theta} = \theta_0$):

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

A confidence *interval* is a one-dimensional confidence region and is obtained when either θ is one-dimensional ($p = 1$) or we are interested in a single component of θ . In the latter case we partition θ as $\theta = [\beta\ \psi]^{\mathsf{T}}$ where β is a scalar (parameter of interest) and ψ is $(p-1)$-dimensional (the nuisance parameters). The maximum likelihood estimate $\hat{\theta}$ is now $[\hat{\beta}\ \hat{\psi}]^{\mathsf{T}}$. The symbols ψ and $\hat{\psi}$ can be ignored if θ is one-dimensional.

The confidence interval will be of form $L \leqslant \beta_0 \leqslant U$ and is given by:

$$\left\{\beta_0 : S(\hat{\beta}, \hat{\psi}) - S(\beta_0, \tilde{\psi}) \leqslant \frac{1}{2} C_{1,1-\alpha}\right\}$$

where $\tilde{\psi}$ is defined by $S(\beta_0, \tilde{\psi}) = \max_{[\beta\ \psi]^{\mathsf{T}} \in \Theta, \beta = \beta_0} S(\beta, \psi)$.

Other Tests Based on the Likelihood

The above tests and confidence regions are based on the difference in log-likelihoods and are referred to as *likelihood-ratio* tests etc. They are (in my view) the best ones to use. For historical and computational reasons two other approaches are commonly seen. They are based on approximating the log-likelihood function by a quadratic and are both asymptotically equivalent to likelihood-ratio methods. The tests are in practice only as good as the quadratic approximation (usually good enough)

I shall present the approximate methods using the simplest case: a hypothesis test for a single scalar parameter (that is: $p = 1$ and $q = 0$). In theoretical work the expectation $\mathcal{E}S''(\theta)$ is often used instead of the observed $S''(\theta)$: this is rarely practicable (and arguably not desirable) in survival analysis.

The *Wald* Test

The log-likelihood is approximated by a quadratic at $\beta = \hat{\beta}$. The statistic for testing the null hypothesis that $\beta = \beta_0$ is

$$-\left(\hat{\beta} - \beta_0\right)^2 S''(\hat{\beta})$$

which is compared with the chisquare(1) distribution. Many computer programs report the reciprocal of the square root of $S''(\hat{\beta})$ as the estimated standard deviation of $\hat{\beta}$ (the 'standard error').

The *Score* Test

The log-likelihood is approximated by a quadratic at $\beta = \beta_0$. This has the huge computational advantage that the log-likelihood does not have to be maximized. The test statistic is:

$$\frac{[S'(\beta_0)]^2}{-S''(\beta_0)}\ , \tag{2}$$

again, compared with the chisquare(1) distribution.

Exercise (hard(ish)): show that the score test applied to a proportional hazards model of a two group comparison gives the log-rank test. Hint: ignore the denominator in both (2) and the log-rank statistic as they merely normalise the variance to unity – concentrate on showing the numerators are proportional.

References

1. Therneau T. M. and Grambsch P. M. (2000) *Modelling Survival Data – Extending the Cox Model.* Springer-Verlag [see chapter three for useful summary and examples and an illustration of what to do when the quadratic approximation breaks down]

2. Wilks S. S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics* 9:60-62