SOR4007 : Survival Analysis

Jean Peyen
Queen's University Belfast

2023/2024

Disclaimer

This document is designed as a concise guide to survival analysis. It is not intended to replace the need for additional reference materials. Some sections are left as exercises to be completed by the reader.

Useful resources

- *General References About Survival Analysis*
  - *Modelling Survival Data in Medical Research*, by David Collett and Alan Kimber. Chapman and Hall/CRC; 3rd edition (2014).
  - *Survival Analysis: A Self-Learning Text*, by David G. Kleinbaum and Mitchel Klein. Springer; 3rd edition (2012).
- *Python and R Tools*
  - *The R Book*, by Michael J. Crawley. Wiley; 2nd edition (2012).
  - *Lifelines: Survival Analysis in Python*, by Cameron Davidson-Pilon. Available as an open-source Python package with ongoing updates.
- *Continuous Time Markov Processes*
  - *Essentials of Stochastic Processes*, by Richard Durrett. Springer; 3rd edition (2016).

Chapter 1

Introduction to survival data

1.1 Features of survival data

Survival data consist in a time origin until the realisation of an event of interest e.g.

1. Origin : a subject is enrolled in a study, Event : death,
2. Origin : beginning of a treatment, Event : patient is cured,
3. Origin : loan start date, Event : repayment,
4. Origin : date of manufacture, Event : failure.

Survival data are often subject to *right-censoring*, that is when in the current set of data the event of interest has not occurred for some subjects e.g.

1. The subject is still alive at the end of the observation period,
2. The subject stopped the treatment (for instance because of side-effects) or left the study,
3. The loan hasn't been re-payed yet,
4. The machine is still functioning or has been stopped.

We can define other forms of censoring that will not be considered in this course such as

- *Left censoring :* the event is known to have occurred before the observation period,
- *Interval censoring :* the exact time of event is unknown but it is known to have occurred during a particular interval of time.

Survival times are *positively skewed* i.e. the histogram will have a longer tail to the right of the interval that contains the largest number of observations.

For each subject i, survival data is generally recorded as a triplet $\{t_i, d_i, \mathbf{x}_i\}$ where

- t_i is the recorded time from time zero to death or censoring for subject i. The terminal event is well defined and need not be actual death.
- d_i is an indicator variable equal to 1 if the subject i died at time t_i and equal to 0 if the subject was censored at that time.

- $\mathbf{x}_i$ contains covariate information about the subject i (e.g. age, sex, length of disease onset, severity of disease, treatment etc.). The covariate information is typically vector-valued.

1.2 Example

Data : leukemia dataset available in the Python package *lifelines*

	t	status	sex	logWBC	Rx
0	35	0	1	1.45	0
1	34	0	1	1.47	0
2	32	0	1	2.20	0
3	32	0	1	2.53	0
4	25	0	1	1.78	0
5	23	1	1	2.57	0
...					
20	6	1	0	3.28	0
21	23	1	1	1.97	1
22	22	1	0	2.73	1
23	17	1	0	2.95	1
24	15	1	0	2.30	1
25	12	1	0	1.50	1
...					

Event of interest : death (status = 1)

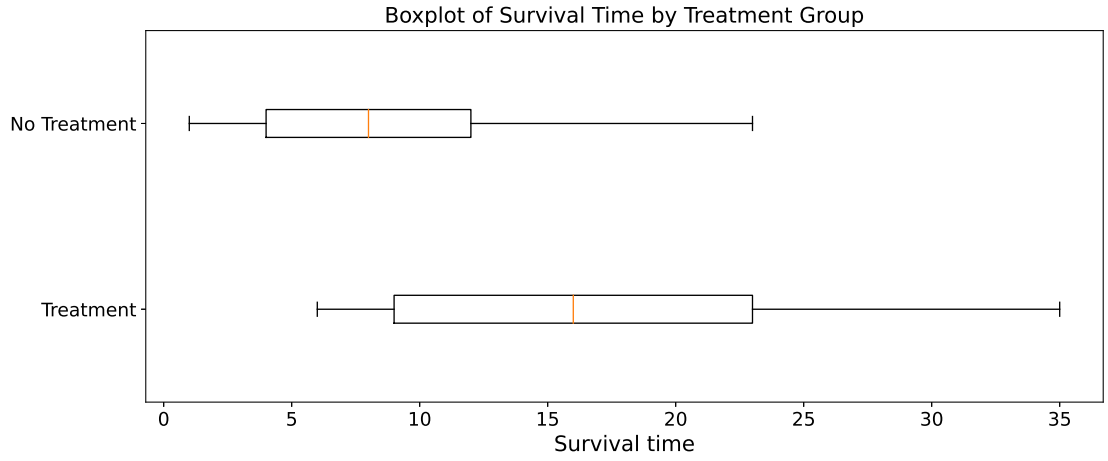
Recorded time : expressed in weeks

Covariates : subjects are characterised by 3 covariates

- Rx : 0 for the treated group, 1 for the control group
- logWBC : logarithm of the white blood cells count
- sex : 0 for females, 1 for males

Here we will just consider the treatment status.

Question Does the treatment increase the survival time ?



Remarks

- t-test cannot be used as the distribution of survival times is skewed (thus it is not Gaussian)
- subjects with status 0 are still alive at the end of the observation period (right-censoring)

1.3 Modelling survival

1.3.1 Density function

A survival time T is interpreted as a continuous *random variable*. Its law can be represented by a *density function* $f(\cdot)$

$$F(t) := P(T \leq t) = \int_0^t f(u)\, \mathrm{d}u \tag{1.1}$$

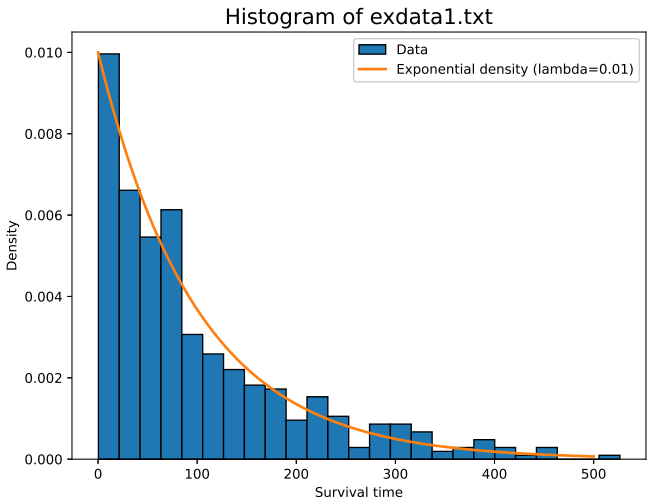
The integral F is called the *cumulative distribution function*.

Example

The exponential distribution is a typical example of law for survival times

$$\begin{aligned} T &\sim \exp(\lambda) \text{ with } \lambda > 0 \\ f(t) &= \lambda e^{-\lambda t} \\ F(t) &= 1 - e^{-\lambda t}. \end{aligned}$$

Let us look at the dataset "exdata1.txt" available on the website. It consists of 496 simulated survival times (no censoring, the sample mean is 101.1894, and the median is 68.12) from an exponential distribution with $\lambda = 0.01$. Here below we give a histogram with the exponential density



If the parameter λ was unknowned, we could check its value from the data using its *maximum likelihood estimation* (MLE).

Here the likelihood and the log-likehood are given by

$$L(\lambda) = \prod_{i=1}^{496} \lambda e^{-\lambda t_i} = \lambda^{496} e^{-\lambda \sum_{i=1}^{496} t_i}$$

$$\ell(\lambda) = \log L(\lambda) = 496 \log \lambda - \lambda \sum_{i=1}^{496} t_i$$

in order to maximise the (log-)likelihood, we need to find a critical point

$$\frac{\partial \ell(\lambda)}{\partial \lambda}\Big|_{\hat{\lambda}} = 0 \Leftrightarrow \frac{496}{\hat{\lambda}} - \sum_{i=1}^{496} t_i = 0$$

here we obtain the estimate

$$\hat{\lambda} \approx 0.0099.$$

1.3.2 Survival function

It may not be easy to see where censored observations fit in with the density function. However, we know that up to the censor time, at least, the censored subject was still alive. It is therefore convenient to think in term of *survival function*

$$S(t) := P(T > t) = 1 - F(t) = 1 - \int_0^t f(u)\,\mathrm{d}u \tag{1.2}$$

1.3.3 Hazard function

The *hazard function* at time t is the instantaneous death rate given that the invidividual survived until then

$$h(t) := \lim_{\delta t \to 0} \frac{P(t \leq T \leq t + \delta t | T \geq t)}{\delta t}. \tag{1.3}$$

While it is not a probability in itself we can observe that, by definition, it estimates the probability for an individual alive at time t to die in the time interval $(t, t+\delta t]$

$$h(t)\delta t \sim P(t \leq T \leq t + \delta t | T \geq t).$$

1.3.4 Relationship between density, survivor, and hazard functions

Here we summarise some useful formulas.

$$h(t) = \frac{f(t)}{S(t)} \tag{1.4}$$

$$f(t) = \frac{-\mathrm{d}S(t)}{\mathrm{d}t} \tag{1.5}$$

$$h(t) = \frac{-\mathrm{d}\log S(t)}{\mathrm{d}t} \tag{1.6}$$

$$S(t) = \exp\left(-\int_0^t h(u)\mathrm{d}u\right) = \exp - H(t), \tag{1.7}$$

where $H(t) := \int_0^t h(u)\mathrm{d}u$ is called the *cumulative hazard function*.

Exercise

Prove these formulas.

Chapter 2

Estimation of the survival function

2.1 Fitting a parametric probability distribution

We have seen in the previous chapter and in the exercises that when a parametric distribution for the survival times is assumed, the model parameters can be estimated using the maximum likelihood or the method of moments. Some commonly used distributions are exponential, Weibull, and log-normal. Parametric models allow for extrapolation beyond the observed data, however the choice of an adequate model requires an understanding of the nature of the data and field knowledge. Assuming that censoring is independent of the survival times (it is said to be *non-informative*), the maximum likelihood method can be adapted to

incorporate right-censored times

$$L(\theta; \{t_i\}) = \prod_{t_i \text{ uncensored}} f(t_i; \theta) \prod_{t_i \text{ censored}} S(t_i; \theta). \tag{2.1}$$

Exercise Prove that the maximum likelihood estimator for the exponential distribution with censored times is given by

$$\hat{\lambda} = \frac{\sharp \text{ uncensored}}{\sum_{i=1}^{k} t_i}. \tag{2.2}$$

Questions

- Why did we assume that the censoring process is non-informative?
- What issues arise if you want to use censored times in the method of moments?

2.2 Empirical survival function

An intuitive way to estimate the survival function $S(t)$ from a sample is to calculate the proportion of elements that exceed t. The validity of this estimation is guaranteed by the following theorem

Theorem (Glivenko–Cantelli)**.** *We assume that* $\{T_1, T_2, \cdots\}$ *is an infinite sequence of iid variables with survival function* S*. We define the* empirical survival function

$$\hat{S}_k(t) := \frac{\sharp\{T_i \,:\, i \leq k,\, T_i > t\}}{k}, \tag{2.3}$$

then $\hat{S}_k$ *converges to* S almost surely *i.e.*

$$\mathbb{P}\left(\lim_{k\to\infty} \sup_t \mid \hat{S}_k(t) - S(t) \mid = 0\right) \to 1. \tag{2.4}$$

This however requires to have a non-censored sample and does not allow to make use of the information from the censored times.

2.3 Life-table/actuarial estimate

We partition the times into disjoint intervals $I_j = [t'_j, t'_{j+1}[$. Let n_j denote the number of individual alive at time t_j, c_j the number of times censored during I_j and d_j the number of deaths during I_j. If we assume that the censoring process occurs uniformly throughout the I_j, the average number of subjects at risk during this interval is

$$n'_j = n_j - \frac{c_j}{2}. \tag{2.5}$$

In I_j, the probability of death is therefore estimated by the number of death divided by the number of subjects at risk that is d_j/n'_j. The probability to survive beyond the interval I_j is the complementary that is $(n'_j - d_j)/n'_j$. Assuming that we have a time $t'_j \leq t < t'_{j+1}$ we observe that the survival function at t may be approximated by

$$S(t) = \mathbb{P}(T > t) \simeq \prod_{j:t_j \leq t} \mathbb{P}\left(T > t'_{j+1} \mid T > t'_j\right). \tag{2.6}$$

Thus we define the *actuarial estimate* of the survival function

$$\hat{S}(t) := \prod_{j:t_j \leq t} \left(\frac{n'_j - d_j}{n'_j}\right). \tag{2.7}$$

Question

What are the benefits and the limitations of this estimation compared to the empirical survival function?

2.4 Kaplan–Meier estimate

In this section we assume that the times are distinct, this happens almost surely if we assume that the time distribution is continuous. Let $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ be **distinct ordered** event times observed in the data. Let $n_{(j)}$ denote the number of subjects at risk until $t_{(j)}$. Assuming that δt is a sufficiently small increment of time, the interval $[t_{(j)} - \delta t, t_{(j)}]$ contains only one event, thus the probability of death during this interval is estimated by $1/n_{(j)}$. If we get back to the times t_j, the probability of death during small intervals $[t_j - \delta t, t_j]$ is d_j/n_j, where d_j indicates if t_j is an event time or a censored time. The probability of survival is its complementary $(n_j - d_j)/n_j$. Since no deaths occur outside of these small intervals, the probability to survive outside of the $[t_{(j)} - \delta t, t_{(j)}]$ is considered to be equal to 1. This gives the *Kaplan–Meier* estimate of the survival function at t

$$\hat{S}(t) := \prod_{j:t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right). \tag{2.8}$$

Question

Why is it called the *product-limit estimate*? What are the benefits and the drawbacks compared to the actuarial estimate?

2.5 Nelson–Aalen estimate

Keeping the notations of the previous section, we recall that the hazard function represents the instantaneous rate of death for subjects that are at risk. Since we only observe deaths at the times $t_{(j)}$ we may heuristically decompose $h(t)$ into Dirac distributions

$$\hat{h} = \sum_{j=1}^{k} \frac{d_j}{n_j} \cdot \delta_{t_j},$$

thus the cumulative hazard rate is estimated by

$$\hat{H}(t) = \int_0^t \hat{h}(u)\, \mathrm{d}u = \sum_{j:t_j \leq t} \frac{d_j}{n_j}. \tag{2.9}$$

Hence the *Nelson–Aalen* estimate of the survival function

$$\hat{S}(t) = e^{-\hat{H}(t)} = \prod_{j:t_j \leq t} e^{-d_j/n_j}. \tag{2.10}$$

Question

With a first order Taylor expansion, prove that the Kaplan–Meier estimate is an approximation of the Nelson–Aalen estimate (you may assume that d_j is small compared to n_j and discuss the validity of this assumption later).

Chapter 3

Confidence intervals for the Kaplan–Meier estimate

So far, we have only computed point estimates of the survival function. The goal of this chapter is to provide insights about the distribution and the variance of the Kaplan–Meier estimate in order to derive confidence intervals. In order to do so, we need to introduce an important result known as the *Delta method*.

3.1 Delta method

Theorem. *Suppose that θ and σ^2 are finite and that $\hat{\theta}_n$ is such that*

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \tag{3.1}$$

then for all C^2 *function* g *such that* $g'(\theta) \neq 0$

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2 \cdot g'(\theta)^2), \tag{3.2}$$

Proof. First of all let us write a Taylor approximation of $g(\hat{\theta}_n)$

$$g(\hat{\theta}_n) = g(\theta) + (\hat{\theta}_n - \theta)g'(\theta) + (\hat{\theta}_n - \theta)^2 O(1).$$

Thus we have

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) = \sqrt{n}\left[(\hat{\theta}_n - \theta)g'(\theta) + (\hat{\theta}_n - \theta)^2 O(1)\right].$$

Subsequently, we can ignore the reminder since $\sqrt{n}(\hat{\theta}_n - \theta)^2$ goes to zero almost surely.

Finally we can prove the theorem with the characteristic function method.

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\left[\exp iz\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right)\right] &= \lim_{n\to\infty} \mathbb{E}\left[\exp i(zg'(\theta))\sqrt{n}\left(\hat{\theta}_n - \theta\right)\right] \\ &= \exp\left(-\frac{(zg'(\theta))^2\sigma^2}{2}\right).\end{aligned}$$

□

3.2 Greenwood's formula

By definition of the Kaplan–Meier estimator

$$\mathrm{Var}[\hat{S}(t)] = \prod_{j:t_j \leq t} \hat{p}_j$$

where

$$\hat{p}_j = \frac{n_j - d_j}{n_j}.$$

It will be more simple to take the logarithm in order to replace the product by a sum. Assuming the independence of the n_j we have

$$\mathrm{Var}[\log \hat{S}(t)] = \sum_{j:t_j \leq t} \mathrm{Var}[\log \hat{p}_j].$$

Thanks to the delta method we can work with $\mathrm{Var}[\hat{p}_j]$ instead of $\mathrm{Var}[\log \hat{p}_j]$.

We may assume that $(n_j - d_j)$ is a binomial random variable with "true" parameters n_j and p_j.

$$\mathrm{Var}[\hat{p}_j] = \frac{p_j(1-p_j)}{n_j}.$$

We may assume that the distribution of $(n_j - d_j)$ is approximated by the normal distribution. By applying the delta method we obtain

$$\mathrm{Var}[\log \hat{p}_j] \simeq \left(\frac{\partial \log p_j}{\partial p_j}\right)^2 \mathrm{Var}[p_j] = \frac{1-p_j}{p_j n_j} \simeq \frac{1-\hat{p}_j}{\hat{p}_j n_j} = \frac{d_j}{n_j(n_j-d_j)},$$

again applying the delta method we have

$$\mathrm{Var}[\hat{S}(t)] \simeq \hat{S}(t)^2 \,\mathrm{Var}[\log \hat{S}(t)] \simeq \hat{S}(t)^2 \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j-d_j)}.$$

taking the square root gives the standard error

$$\mathrm{se}[\hat{S}(t)] \simeq \hat{S}(t) \sqrt{\sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j-d_j)}}. \tag{3.3}$$

This is the *Greenwood formula*.

Question

Can you justify the assumptions made in this proof?

3.3 Confidence intervals

So far we have only used the delta method in order to estimate a variance of the Kaplan–Meier estimator. But we may also use the fact that it is approximately Gaussian in order to obtain confidence intervals for the survival function.

Exercise

Derive the symmetric confidence interval for $S(t)$, with level $1 - p$

$$\left[\max\left\{\hat{S}(t) - Q_{\alpha/2}\,\mathrm{se}[\hat{S}(t)], 0\right\}, \min\left\{\hat{S}(t) + Q_{\alpha/2}\,\mathrm{se}[\hat{S}(t)], 1\right\}\right], \tag{3.4}$$

where $Q_{p/2}$ is the value corresponding to the $1 - p/2$ percentile of the standard normal distribution. Add the 95% confidence interval to your Python code for the Kaplan–Meier estimate.

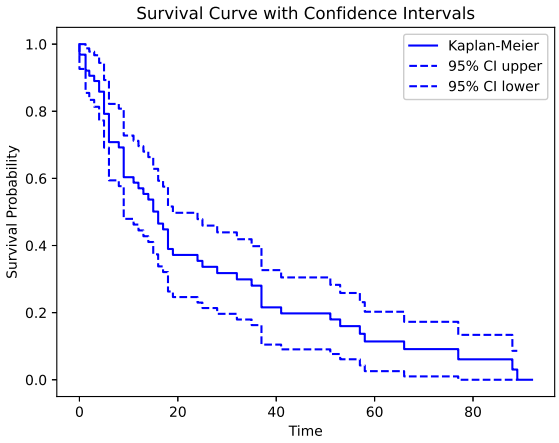


Table 3.1: Standard normal distribution table

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Chapter 4

Comparing groups

We consider a total population with n subjects split into two groups of size n_1 and n_2. Time periods $1, 2, \cdot, J$ are delineated by events (deaths and censoring) occurring in either group. We denote by n_{1j} (resp. n_{2j}) the number of subjects at risk in the j-th period for the group 1 (resp. group 2) and d_{1j} (resp. d_{2j}).

The goal of this chapter is to test the hypothesis H_0 that both groups are identical. If this hypothesis were valid, assuming that the group size is large enough, we should expect the number of deaths d_{1j} to be close to the following

$$e_{1j} := n_{1j}\frac{d_j}{n_j}. \tag{4.1}$$

In general, if H_0 is satisfied, d_{1j} follows the hypergeometric distribution with parameter n_j, d_j and n_{1j}. If the sample size is large, this distribution is well approximated by the normal distribution, thus it is possible to test if the d_{1j} have a statistically significant difference from the e_{1j} with a χ^2 test.

4.1 Log-rank test

The log-rank measures the total difference between the observed numbers of deaths in each period and their expectation under H_0

$$U_L := \sum_{j=1}^{J} d_{1j} - e_{1j}. \tag{4.2}$$

Under H_0 the variance of this statistic is given by

$$V_L := \sum_{j=1}^{J} v_{1j}, \qquad v_{1j} := \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}. \tag{4.3}$$

In which case the distribution χ_1^2 is used to test if the value of U_L^2 / V_L is significant. This test is used when the hazards are proportional (we will talk more about this assumption later).

4.2 Wilcoxon test

The Wilcoxon test is similar but it applies weightings to the observations

$$U_W := \sum_{j=1}^{J} n_j (d_{1j} - e_{1j}). \tag{4.4}$$

$$V_W := \sum_{j=1}^{J} n_j^2 v_{1j}. \tag{4.5}$$

Table 4.1: χ^2 distribution table

d	0.995	0.99	0.975	0.95	0.9	0.5	0.1	0.05	0.025	0.01	0.005
1	7.8794	6.6349	5.0239	3.8415	2.7055	0.4549	0.0158	0.0039	0.0010	0.0002	0.0000
2	10.5966	9.2103	7.3778	5.9915	4.6052	1.3863	0.2107	0.1026	0.0506	0.0201	0.0100
3	12.8382	11.3449	9.3484	7.8147	6.2514	2.3660	0.5844	0.3518	0.2158	0.1148	0.0717
4	14.8603	13.2767	11.1433	9.4877	7.7794	3.3567	1.0636	0.7107	0.4844	0.2971	0.2070
5	16.7496	15.0863	12.8325	11.0705	9.2364	4.3515	1.6103	1.1455	0.8312	0.5543	0.4117
6	18.5476	16.8119	14.4494	12.5916	10.6446	5.3481	2.2041	1.6354	1.2373	0.8721	0.6757
7	20.2777	18.4753	16.0128	14.0671	12.0170	6.3458	2.8331	2.1673	1.6899	1.2390	0.9893
8	21.9549	20.0902	17.5345	15.5073	13.3616	7.3441	3.4895	2.7326	2.1797	1.6465	1.3444
9	23.5893	21.6660	19.0228	16.9190	14.6837	8.3428	4.1682	3.3251	2.7004	2.0879	1.7349
10	25.1882	23.2093	20.4832	18.3070	15.9872	9.3418	4.8652	3.9403	3.2470	2.5582	2.1559
11	26.7568	24.7250	21.9200	19.6751	17.2750	10.3410	5.5778	4.5748	3.8157	3.0535	2.6032
12	28.2995	26.2170	23.3367	21.0261	18.5493	11.3403	6.3038	5.2260	4.4038	3.5706	3.0738
13	29.8195	27.6882	24.7356	22.3620	19.8119	12.3398	7.0415	5.8919	5.0088	4.1069	3.5650
14	31.3193	29.1412	26.1189	23.6848	21.0641	13.3393	7.7895	6.5706	5.6287	4.6604	4.0747
15	32.8013	30.5779	27.4884	24.9958	22.3071	14.3389	8.5468	7.2609	6.2621	5.2293	4.6009
16	34.2672	31.9999	28.8454	26.2962	23.5418	15.3385	9.3122	7.9616	6.9077	5.8122	5.1422
17	35.7185	33.4087	30.1910	27.5871	24.7690	16.3382	10.0852	8.6718	7.5642	6.4078	5.6972
18	37.1565	34.8053	31.5264	28.8693	25.9894	17.3379	10.8649	9.3905	8.2307	7.0149	6.2648
19	38.5823	36.1909	32.8523	30.1435	27.2036	18.3376	11.6509	10.1170	8.9065	7.6327	6.8439
20	39.9968	37.5662	34.1696	31.4104	28.4120	19.3372	12.4426	10.8508	9.5908	8.2604	7.4338
21	41.4011	38.9322	35.4789	32.6706	29.6151	20.3369	13.2396	11.5913	10.2829	8.8972	8.0337
22	42.7957	40.2894	36.7807	33.9245	30.8133	21.3366	14.0415	12.3380	10.9824	9.5425	8.6426
23	44.1813	41.6384	38.0756	35.1725	32.0069	22.3362	14.8478	13.0905	11.6890	10.1957	9.2604
24	45.5585	42.9798	39.3641	36.4150	33.1962	23.3359	15.6581	13.8488	12.4015	10.8564	9.8862
25	46.9279	44.3140	40.6465	37.6525	34.3816	24.3356	16.4719	14.6119	13.1197	11.5240	10.5197
26	48.2899	45.6417	41.9231	38.8855	35.5632	25.3353	17.2893	15.3798	13.8434	12.1981	11.1602
27	49.6448	46.9628	43.1942	40.1139	36.7410	26.3349	18.1100	16.1519	14.5730	12.8785	11.8076
28	50.9929	48.2780	44.4600	41.3379	37.9152	27.3346	18.9338	16.9279	15.3076	13.5647	12.4615
29	52.3344	49.5874	45.7206	42.5576	39.0857	28.3342	19.7608	17.7075	16.0475	14.2564	13.1215
30	53.6696	50.8920	46.9767	43.7732	40.2531	29.3339	20.5908	18.4908	16.7919	14.9535	13.7872

Chapter 5

Parametric proportional hazards models

So far we only made inference based on the observed times. Now we want to be able to establish a relationship between the characteristics of the subjects (the covariates) and the hazards.

In this chapter and the next we will assume that the hazards between subjects are proportional. The general proportional hazards assumption consists in stating that the hazard rate for a subject i is a separable function of the time t and the covariates characterising the subject $\mathbf{x}_i$

$$h_i(t) = \underbrace{h_0(t)}_{\text{baseline hazard}} \times \underbrace{\psi(\mathbf{x}_i)}_{\text{relative hazard}} . \tag{5.1}$$

In this chapter we focus on models that take the following form

$$h_i(t) = \underbrace{h_0(t;\theta)}_{\text{baseline hazard}} \times \underbrace{\exp(\beta \cdot \mathbf{x}_i)}_{\text{relative hazard}}, \tag{5.2}$$

where the parameter θ of the baseline and the regression coefficients β need to be fitted. This can be done by maximising the (log-)likelihood method.

Exercise

Calculate the log-likelihood.

5.1 Newton–Raphson algorithm

For convenience let us define $\Theta := (\theta, \beta)$. Maximising the log-likelihood is equivalent to finding a zero of the Jacobian of the log-likelihood. The Newton–Raphson algorithm is an itervative method that consist in using the current parameter estimate $\hat{\Theta}_m$ to compute the next estimate $\hat{\Theta}_{m+1}$. In general, a first order Taylor expansion gives us the following equation

$$J(\hat{\Theta}_{m+1}) = J(\hat{\Theta}_m) + Hess(\hat{\Theta}_m) \cdot (\hat{\Theta}_{m+1} - \hat{\Theta}_m) + o(\hat{\Theta}_{m+1} - \hat{\Theta}_m). \tag{5.3}$$

Where J is the Jacobian and $Hess$ is the Hessian of the log-likelihood. In order to estimate critical points, we may find a zero for this quantity assuming that the reminder is null. This means that the next estimate satisfies the following equation

$$0 = J(\hat{\Theta}_m) + Hess(\hat{\Theta}_m) \cdot (\hat{\Theta}_{m+1} - \hat{\Theta}_m), \tag{5.4}$$

or more explicitly

$$\hat{\Theta}_{m+1} = \hat{\Theta}_m - Hess(\hat{\Theta}_m)^{-1} \cdot J(\hat{\Theta}_m). \tag{5.5}$$

5.2 Weibull model

In the Weibull model the hazard rate is given by

$$h_i(t) = (\lambda\gamma t^{\gamma-1}) \times \exp(\beta \cdot \mathbf{x}_i). \tag{5.6}$$

Where λ is called the scale parameter and γ is called the shape parameter. In some formulations the parameter may be put inside the exponential as an "intercept" coefficient

$$h_i(t) = (\gamma t^{\gamma-1}) \times \exp(\beta_0 + \beta \cdot \mathbf{x}_i). \tag{5.7}$$

This model is suitable when the hazard rate is a monotonic function of t.

Exercise

Calculate the log-likelihood, the Jacobian and the Hessian for the Weibull model.

Chapter 6

Cox model

Fully parametric models are convenient as they can be used to extrapolate times, however the choice of a baseline requires prior knowledge about the features of the survival times distribution. The Cox model aims to tackle this issue by only focusing on the relative hazards. It allows to understand the effect of the covariates on the survival times.

6.1 The Cox proportional hazards model

As in the previous chapter, we assume that the hazard rates take the following form

$$h_i(t) = \underbrace{h_0(t)}_{\text{baseline hazard}} \times \underbrace{\exp(\beta \cdot \mathbf{x}_i)}_{\text{relative hazard}} . \tag{6.1}$$

The Cox model is fitted by maximising the partial likelihood

$$L(\beta) := \prod_{i=1}^{n} \left(\frac{\exp(\beta \cdot \mathbf{x}_i)}{\sum_{j:t_j \geq t_i} \exp(\beta \cdot \mathbf{x}_j)} \right)^{d_i} . \tag{6.2}$$

This partial likelihood consists in measuring for each death time, t_i such that $d_i = 1$, the probability that it is actually the death of the subject i, conditionally on everything that happened until then. In practice, it can be fitted with the Newton–Raphson algorithm.

Exercise

Calculate the log-likelihood, its Jacobian and its Hessian.

Remark

Some conventions may include the scale factor as an intercept coefficient in the relative hazard part instead of the baseline

$$h_i(t) = \underbrace{h_0(t)}_{\text{baseline hazard}} \times \underbrace{\exp(\beta_0 + \beta \cdot \mathbf{x}_i)}_{\text{relative hazard}} . \tag{6.3}$$

6.2 Confidence intervals and likelihood ratios

Once the regression coefficients are fitted the corresponding covariance matrix is given by

$$\mathrm{Cov}(\hat{\beta}) = -Hess^{-1}(\hat{\theta}). \tag{6.4}$$

The regression coefficients are approximately normally distributed. In particular it allows to give (asymptotic) confidence interval for the hazard ratios.

Exercise

1. We consider two subjects labelled 1 and 2, with covariates $\mathbf{x}_1$ and $\mathbf{x}_2$. Show that a (asymptotic) confidence interval of probability p for the hazard ratio h_1/h_2 is given by

$$\left[\exp\left(\hat{\beta}\cdot(\mathbf{x}_1-\mathbf{x}_2)\pm Q_{1-p/2}\sqrt{(\mathbf{x}_1-\mathbf{x}_2)^t(-Hess^{-1})(\mathbf{x}_1-\mathbf{x}_2)}\right)\right], \quad (6.5)$$

   where $Q_{1-p/2}$ is the $1-p/2$ quantile of the standard normal distribution.

2. The survival times of 227 lung cancer patients have been recorded, along with their age (centered on 62), their gender and their ECOG score (it measures the level of disability, a score of 0 corresponds to a fully able person). A Cox proportional hazards model has been fitted to these data. The fitted coefficient and the covariance matrix are given here below

```
             coef  exp(coef)   se(coef)

age          0.01       1.01       0.01
sex         -0.55       0.58       0.17
ph.ecog      0.46       1.59       0.11

         age       sex   ph.ecog

age         0.000086  0.000004 -0.000182
sex         0.000004  0.028136 -0.000930
ph.ecog    -0.000182 -0.000930  0.012900
```

Calculate 95% confidence intervals for the coefficients and for the hazard ratios between a 62 years old female and a 50 years old male with an identical ECOG score.

6.3 Estimation of the baseline

One of the defining features of the Cox-model is that it is a semi-parametric model which allows to assess the effect of the covariates without the need of a baseline. If it is needed to provide a baseline, this can be done with the Breslow estimate

$$\hat{h}_0 := \sum_{j=1}^{n} \frac{d_j}{\sum_{\ell:t_\ell \le t_j} \exp(\hat{\beta} \cdot \mathbf{x}_\ell)} \cdot \delta_{t_j}. \tag{6.6}$$

Exercise

Check that it is an extension of the Nelson–Aalen estimate and calculate $\hat{H}_0(t)$.

6.4 Evaluating a Cox model

6.4.1 Likelihood ratio

We consider a Cox model A as a nested model within the more comlex model B, where all covariates included in model A are also present in model B. Their performance can be compared by testing the ratio of their likelihood. Under the

hypothesis H_0 that the (simpler) model A is true the following quantity

$$-2 \log \left(\frac{L_A}{L_B} \right) \tag{6.7}$$

(approximately) follows the χ^2_d distribution where d is the difference between the number of parameters in model A and model B.

Question

How would you compare a Cox model with a baseline model without any covariates?

6.4.2 Cox–Snell residuals

Exercise

A survival model has been fitted to some data. We define the Cox-Snell residual for a subject i as

$$r_i := \hat{H}_i(t_i) \tag{6.8}$$

where $\hat{H}_i$ is the estimated cumulative hazards for the subject and t_i is the observed time. Demonstrate that if model fits the data perfectly, the Cox-Snell residuals should follow the standard exponential distribution. Deduce a method to check if a survival model is a good fit for the data.

6.5 Stratified Cox model

When a small number of categorical covariates break the proportional hazards assumption, the data can be separated into multiple strata. The baseline is then allowed to vary between the strata while the set of covariates remain the same in each stratum

$$\underbrace{h_{ij}(t)}_{\text{subject } i \text{ in stratum } j} := \underbrace{h_{0j}(t)}_{\text{baseline in stratum } j} \times \exp(\beta \cdot \mathbf{x}_{ij}). \tag{6.9}$$

The partial likelihood is given by

$$L(\beta) := \prod_{j \text{ stratum}} \prod_{i \text{ subject}} \left(\frac{\exp(\beta \cdot \mathbf{x}_{ij})}{\sum_{\ell : t_\ell \geq t_{ij}} \exp(\beta \cdot \mathbf{x}_{\ell j})} \right)^{d_{ij}}. \tag{6.10}$$

6.6 Time-varying covariates

Studies may involve multiple data entries per subject recording time-varying covariates. Such covariates can be separated into two classes:

- Internal : these covariates are affected by the survival and the progression of the subject. Their value may be unknown at the event times (e.g. blood tests). It may be necessary to extrapolate missing values via diverse methods such as carrying forward the last known value or by interpolation.
- Exteral : the evolution of these covariates can be predicted without following the subject (e.g. age).

The Cox model can be generalised to incorporate time-varying covariates

$$h_i(t) = \underbrace{h_0(t)}_{\text{baseline hazard}} \times \underbrace{\exp(\beta \cdot \mathbf{x}_i(t))}_{\text{relative hazard}}. \tag{6.11}$$

In this model, the hazards ratios are not constant anymore unless comparing two subjects such that the vector $\mathbf{x}_i - \mathbf{x}_j$ is constant. The model can be fitted by maximising a partial (log-)likelihood that incorporates the variation of the covariates

$$L(\beta) := \prod_{i=1}^{n} \left(\frac{\exp(\beta \cdot \mathbf{x_i}(t_i))}{\sum_{\ell : t_\ell \geq t_i} \exp(\beta \cdot \mathbf{x}_\ell(t_i))} \right)^{d_i} \tag{6.12}$$

Exercise

This exercise focuses on the Stanford heart transplants dataset available in the lifelines package.

1. Fit a Cox model incorporating a constant indicator denoting whether subjects have received a heart transplant at any point during the observation period. Compare the hazard rates between subjects who have ever received a transplant and subjects who never received a transplant.

2. Fit a model where the transplant status is a time-varying indicator.

3. Discuss the concept of immortal bias.

Remark

Adding an external time-varying covariate that records time (or a function of time) to a Cox proportional hazards model is a way to assess whether the proportional hazards assumption is met. If the coefficient corresponding to this covariate is statistically different from zero indicates a dynamic that the proportional hazards Cox model cannot capture.

Chapter 7

Accelerated failure time model

Let us start with a formal example where we have two groups. A *treatment* (T) group and a *control* (C) group. The effect of the treatment is to act multiplicatively on the time via an acceleration factor ϕ

$$S_T(t) = S_C(\phi t). \tag{7.1}$$

Exercise

Check the following formulas

$$H_T(t) = H_C(\phi t), \quad h_T(t) = \phi h_T(\phi t), \quad f_T(t) = \phi f_T(\phi t). \tag{7.2}$$

The accelerated failure time model (AFT) is a general model, where the

acceleration factor is determined by the covariates. For a subject i with covariates vector $\mathbf{x}_i$, the hazard rate is given by

$$H_i(t) := H_0(\exp(\beta \cdot \mathbf{x}_i)\ t). \tag{7.3}$$

The baseline is assumed to be in a parametric family. This model can be fitted by maximising the (log-)likelihood as in Section 2.1

$$L(\beta, \sigma) = \prod_i (f_i(t_i))^{d_i} S_i(t_i)^{1-d_i}. \tag{7.4}$$

Weibull baseline

$$h_0(t) = \lambda \gamma t^{\gamma - 1} \tag{7.5}$$

Log-logistic baseline

$$h_0(t) = \frac{\exp(\theta)\kappa t^{\kappa - 1}}{1 + \exp(\theta) t^{\kappa}} \tag{7.6}$$

Exercise

1. Calculate the cumulative hazard, the survival function and the density for the Weibull and the log-logistic AFT.

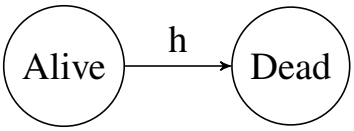
2. Check that the proportional hazards assumption is satisfied by the Weibull AFT.

3. Discuss the features of these two models.

Chapter 8

Multistate models

So far we considered situations where subject could only transition from an 'Alive' state to a 'Dead' state at a variable rate $h(t)$.



Multistate models allow to model more complex pathways

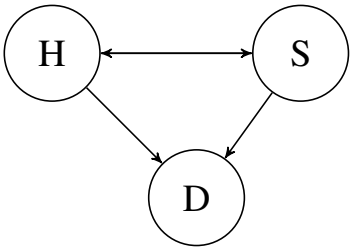


Figure 8.1: Disease model with 3 states

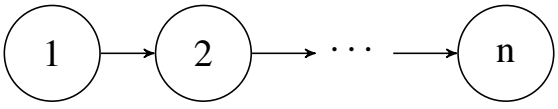


Figure 8.2: Manufacturing chain

This chapter focuses on continuous time, homogeneous Markov models.

8.1 Homogeneous Markov processes

We consider a process $X : \mathbb{R}^+ \to \Omega$. The configuration space Ω is assumed to be finite and the law of the process is given by the Kolmogorov equation

$$\frac{d}{dt}\mathbb{E}[f(X(t)) \mid X(0) = i] = \sum_{j\in\Omega} \lambda_{ij}(f(j) - f(i)), \tag{8.1}$$

where $f : \Omega \to \mathbb{R}$ and the transition matrix $\Lambda = (\lambda_{ij})$ statisfies the following

$$\lambda_{ii} = -\sum_{j\neq i} \lambda_{ij}, \quad \lambda_{ij} \geq 0 \text{ for } j \neq i. \tag{8.2}$$

Exercise

In this exercise we assume that the law at time 0 is known. The law at any time is represented by a vector $\mathbb{P}[X(t)] := (\mathbb{P}[X(t) = i],\ i \in \Omega)$.

1. Prove the following equation

$$\frac{d}{dt}\mathbb{P}[X(t)] = \mathbb{P}[X(0)] \cdot \Lambda. \tag{8.3}$$

   Hint. Use the Kolmogorov equation with indicator functions $f = \mathbf{1}_j$. Then use the total expectation rule.

2. Integrate this equation and deduce the law $\mathbb{P}[X(t)]$.

3. Assume that the system is in a state i at time 0. Prove that the time until the system leaves this state follows the exponential distribution with parameter $\sum_{j\neq i} \lambda_{ij}$.

   Hint. You just need to consider a suitable survival process.

8.2 First-reaction Gillespie algorithm

Exercise

1. Let us consider a set of independent exponential random variables $T_1, \cdots, T_n$ with respective rates $\lambda_1, \cdots, \lambda_n$. Prove that $T := \min_k T_k$ follows the exponential distribution with rate $\sum_k \lambda_k$.

2. Prove the following equation

$$\mathbb{P}[T = t_k] = \frac{\lambda_k}{\sum_\ell \lambda_\ell}. \tag{8.4}$$

3. Use this and the question 3 of the previous section to write an algorithm to simulate the evolution of a homogeneous Markov process.

8.3 Likelihood

Let us consider a subject i experiencing a sequence of transitions at times (t_k), where $t_0 = 0$. Between the times t_k and t_{k+1}, the subject is in the state i_k. Assuming that we have a model entirely determined by a set of parameters σ and that the observations are not censored (i.e. the subject is observed until an absorbing state is reached), the likelihood function for the subject i is given by

$$L_i(\sigma) := \prod_{k=0}^{n-1} \lambda_{i_k i_{k+1}} \exp\left(\lambda_{i_k i_k}(t_{k+1} - t_k)\right), \tag{8.5}$$

where the transition matrix is a function of σ.

Right-censoring can be handled with the following likelihood

$$L_i(\sigma) := \left[\prod_{k=0}^{n-2} \lambda_{i_k i_{k+1}} \exp\left(\lambda_{i_k i_k}(t_{k+1} - t_k)\right)\right] \exp\left(\lambda_{i_{n-1} i_{n-1}}(t_n - t_{n-1})\right). \quad (8.6)$$

The full likelihood is just the product of the individual likelihoods

$$L(\sigma) := \prod_{i \text{ subject}} L_i(\sigma). \quad (8.7)$$