Survival analysis

Christiana Kartsonaki

Abstract

Survival analysis is the analysis of data involving times to some event of interest. The distinguishing features of survival, or time-to-event, data and the objectives of survival analysis are described. Some fundamental concepts of survival analysis are introduced and commonly used methods of analysis are described.

Keywords Cox proportional hazards model; failure times; hazard; Kaplan–Meier curve; survival data; time-to-event data

Introduction

Survival analysis is the analysis of time-to-event data. Such data describe the length of time from a time origin to an endpoint of interest. For example, individuals might be followed from birth to the onset of some disease, or the survival time after the diagnosis of some disease might be studied. Survival analysis methods are usually used to analyse data collected prospectively in time, such as data from a prospective cohort study or data collected for a clinical trial.

The time origin must be specified such that individuals are as much as possible on an equal footing. For example if the survival time of patients with a particular type of cancer is being studied, the time origin could be chosen to be the time point of diagnosis of that type of cancer. Equally importantly, the endpoint or event of interest should be appropriately specified, such that the times considered are well-defined. In the above example, this could be death due to the cancer studied. Then the length of time from the time origin to the endpoint could be calculated.

One of the reasons why survival analysis requires 'special' techniques is the possibility of not observing the event of interest for some individuals. For example individuals may drop out of a study, or they might have a different event, such as in the above example death due to an accident, which is not part of the endpoint of interest. Another possibility is that there might be a time point at which the study finishes and thus if any individuals have not had their event yet, their event time will not have been observed. These incomplete observations cannot be ignored, but need to be handled differently. This is called *censoring*. Another feature of survival data is that distributions are often skewed (asymmetric) and thus simple techniques based on the normal distribution cannot be directly used.

The objectives of survival analysis include the analysis of patterns of event times, the comparison of distributions of survival times in different groups of individuals and examining whether and by how much some factors affect the risk of an event of interest.

Censoring

The most commonly encountered type of censoring and easiest to handle in the analysis is *right censoring*. Right censoring occurs when an individual is followed up from a time origin t_0 up to some later time point t_C and he/she has not had the event of interest, such that all we know is that their event has not occurred up to their censoring time t_C. This may occur, for example, if an individual drops out of a study before the event of interest occurs. Commonly studies are terminated at some specified time and at the end of the study some individuals have not yet had their event. This is sometimes referred to as *administrative* censoring. In some studies the majority of participants are censored. Event and censoring times of 10 patients are illustrated in Figure 1.

Another type of censoring is *left censoring*. Left censoring is the situation in which an individual is known to have had the event before a specific time, but that could be any time before the censoring time. It is also possible to have *interval censoring* where an individual is only known to have had the event between two time points but the exact time of event is not observed.

A different concept is *truncation*. Truncation is something that happens by design. *Left truncation* is the most commonly encountered type of truncation, where individuals enter the study after they have their truncation event (which is not the same as the event being studied). Delayed entry where for example a set of adults are recruited into a study but those who had the event before adulthood are not included at all is very common. *Right truncation* occurs when the entire study population has already experienced the event of interest.

For the standard methods of analysis that we focus on here censoring should be *non-informative*, that is, the time of censoring should be independent of the event time that would have otherwise been observed, given any explanatory variables included in the analysis, otherwise inference will be biased.

An example of informative censoring which must not be ignored is as follows: in a study of survival after a disease diagnosis, patients might be lost to follow up because their condition has become worse and are no longer able to attend appointments. Or in a study of treatments for a non-life-threatening condition, some patients might drop out of the study because their condition has improved and they choose to discontinue treatment. It is usually not possible to know whether the censoring in a study really is non-informative.

Example. Data from a clinical trial on colon cancer adjuvant therapy[1] are used as an illustration. A group of colon cancer patients are followed up from diagnosis to death. That is, the *time scale* has origin the time of diagnosis of colon cancer and endpoint the time of death from colon cancer. The dataset, freely available in the statistical software R[2] (dataset 'colon' in package 'survival'[3]), contains observations on 929 colon cancer patients. These are the first 10 observations on a subset of the variables:

Christiana Kartsonaki ***DPhil*** *Nuffield Department of Population Health, University of Oxford, Oxford, UK. Conflict of interest statement: none.*

id	status	time	sex	age	nodes	differ	surg	node4
1	1	1521	1	43	5	2	0	1
2	0	3087	1	63	1	2	0	0
3	1	963	0	71	7	2	0	1
4	1	293	0	66	6	2	1	1
5	1	659	1	69	22	2	1	1
6	1	1767	0	57	9	2	0	1
7	1	420	1	77	5	2	1	1
8	0	3192	1	54	1	2	0	0
9	0	3173	1	46	2	2	0	0
10	0	3308	0	68	1	2	1	0

The variable ‘status’ indicates whether a patient has died, taking the value 1 if a patient has died and 0 otherwise, and ‘time’ is the survival time since diagnosis in days. ‘age’ is the patients’ age at the time of entry into the study, ‘nodes’ is the number of lymph nodes with detectable cancer and ‘node4’ is a binary variable taking the value 1 if the patient has more than four lymph nodes with cancer and 0 if the patient has fewer than or equal to four positive lymph nodes. The event and censoring times are illustrated in Figure 1.

Some definitions

Let $T \geq 0$ be a random variable representing the survival (or event) time. The *survival* (or *survivor*) *function* is the probability that an individual survives beyond time t,

$$S(t) = \mathbb{P}(T > t), \ \ 0 < t < \infty.$$

The *probability density function* $f(t)$ is the frequency of events per unit time. The probability density function is related to the survival function,

$$f(t) = -\frac{dS(t)}{dt}.$$

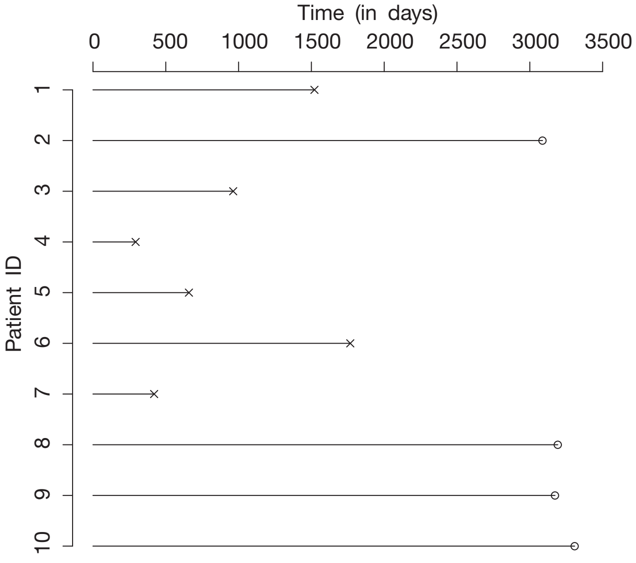

Figure 1 Survival times (×) and censoring times (○) in days.

The *hazard function* is the instantaneous rate at which events occur for individuals which are surviving at time t,

$$h(t) = \lim_{\delta t \to 0^+} \frac{\mathbb{P}(t \leq T < t + \delta t | T \geq t)}{\delta t}$$

and the *cumulative hazard function* is

$$H(t) = \int_0^t h(u)du.$$

The cumulative hazard function is related to the survival function as follows:

$$S(t) = e^{-H(t)}.$$

That is, the higher the hazard, the lower the survival.

Let δ_i be equal to 1 for individual i if individual i had the event and 0 if individual i was censored. Then for a set of possibly right-censored data, the data for individual i can be represented as (t_i, δ_i, x_i), where t_i is the time of event or censoring, δ_i is a censoring indicator and x_i are the covariates, that is, a set of variables representing any other information on that individual.

Then the *likelihood function* is

$$L = \prod_{j \text{ had event}} f(t_j) \prod_{k \text{ censored}} S(t_k) = \prod_{i=1}^{N} h(t_i)^{\delta_i} S(t_i).$$

That is, each individual with an observed event time t_i contributes the hazard rate at t_i multiplied by the survival to t_i and each individual that is censored at t_i contributes the survival to t_i.

Estimation

One objective of the analysis of time-to-event data is given a set of data to estimate and plot the survival function.

A very widely used method of doing that is calculating and plotting a Kaplan–Meier curve. This is a non-parametric method of estimating the survival function. Non-parametric methods are rather simple methods which do not make any distributional assumptions, in this context about the distribution of survival times observed in a study. Non-parametric methods are very useful for summarizing survival data and making simple comparisons but cannot so easily deal with more complex situations.

Let $t_1 < t_2 < \ldots < t_k$ be the observed event times and $n = n_0$ the sample size. Let d_j be the number of individuals who have an event at time t_j, where $j = 1, \ldots, k$, and m_j the number of individuals censored in the interval $[t_j, t_{j+1})$. Then $n_j = (m_j + d_j) + \ldots + (m_k + d_k)$ is the number of individuals at risk just prior to t_j.

The *Kaplan–Meier* (or *product-limit*) *estimator*[4] is a non-parametric estimator of the survival function,

$$\widehat{S}(t) = \prod_{j:\, t_j \leq t} \frac{n_j - d_j}{n_j}.$$

Standard errors can be calculated using *Greenwood’s formula*,[5] which approximates the variance as

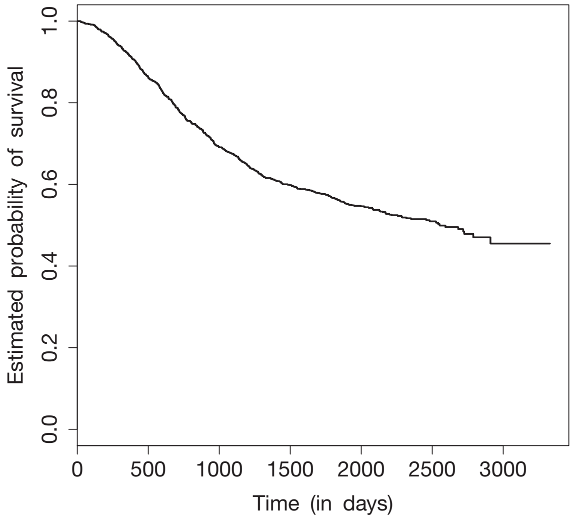

Figure 2 Kaplan–Meier curve for colon cancer data.

$$\widehat{\text{var}}\left\{\hat{S}(t)\right\} = \left\{\hat{S}(t)\right\}^2 \prod_{j:\ t_j \le t} \frac{d_j}{n_j(n_j - d_j)}.$$

Figure 2 is an example of a Kaplan–Meier curve, calculated from the data in the example used above. Confidence intervals can be plotted around the curve. An alternative, less commonly used but very similar, non-parametric estimate of the survival function is the *life table estimator*, based on dividing the time scale into cells.

Kaplan–Meier curves can be used in simple analyses of which the aim is to compare survival times of two or more generally a small number of groups. For example in a clinical trial the researchers might want to look at the survival times of individuals allocated to treatment A and of those allocated to treatment B. In an epidemiological prospective cohort study the researchers might want to contrast the survival times of people who drink alcohol to those who do not. This can be examined by plotting two Kaplan–Meier curves, one for treatment A and one for treatment B in the first example, or one for alcohol drinkers and one for never-drinkers in the second example. In the colon cancer data illustrated above, one might want to compare the survival times of colon cancer patients with up to four positive lymph nodes to those of patients with more than four positive lymph nodes, in order to determine whether having 'many' lymph nodes with cancer is linked to shorter survival after diagnosis. Figure 3 shows the two curves. Note that each event time appears as a 'jump' on a Kaplan–Meier curve. Censoring times are also commonly plotted on a Kaplan–Meier curve, to visualise the amount and patterns of censoring with time.

The Kaplan–Meier curves for the two groups in Figure 3 suggest that at any given time point, a smaller proportion of people with more than four lymph nodes survive beyond that point, compared to those with up to four positive lymph nodes.

To estimate and plot the cumulative hazard function, the *Nelson–Aalen estimator* can be used. The Nelson–Aalen estimator is a non-parametric estimator of the cumulative hazard function,

$$\hat{H}(t) = \sum_{j:\ t_j \le t} \frac{d_j}{n_j} = \sum_{j:\ t_j \le t} \hat{h}_j,$$

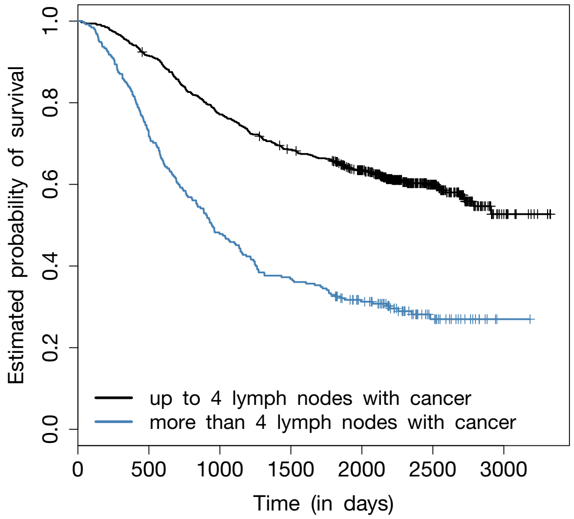

Figure 3 Kaplan–Meier curves for colon cancer patients with up to four lymph nodes with detectable cancer (black) and more than four lymph nodes with cancer (blue). + indicates censoring.

where d_j is the number of individuals who have an event at time t_j, where $j = 1, \ldots, k$, and n_j is the number of individuals at risk just prior to t_j. A very similar alternative is to calculate the Kaplan–Meier estimate of the survival function and take minus its logarithm as an estimate of the cumulative hazard, derived by the relationship between the survival and cumulative hazard functions.

Comparison of survival curves

Another possible objective of the analysis of survival data may be to compare the survival times of two or more groups. A simple test of statistical significance is the *log-rank* or *Mantel–Haenzel test*. It can be used to test whether the survival of individuals in two or more groups is significantly different and it is similar to the χ^2 (chi-squared) test for association. More formally, it tests the hypothesis that survival functions $S_0(t), \ldots, S_p(t)$ are equal, based on samples from each of $p+1$ populations. If h_j denotes the hazard (that is, the conditional failure probability) at time t_j, the null hypothesis associated with the log-rank test is that h_j is common for all $p+1$ samples. The log-rank test statistic compares the observed with the 'expected' number of failures and has an asymptotic χ^2 distribution under the null hypothesis. The degrees of freedom are p (the number of groups minus 1).

Example. In the above example, suppose that we want to compare the survival times of male and female colon cancer patients. Using the log-rank test on these data gives a p-value of 0.89. Thus we do not reject the null hypothesis, that is, we conclude that there is no evidence from these data that the survival times of males are different from those of females.

Parametric models

An alternative basis for estimation and testing in survival analysis is the use of *parametric models*. Parametric methods are methods in which we make assumptions about the patterns of survival times. The distribution of survival times can be represented using continuous parametric survival models. This can be most easily thought of as assuming that the hazard, as a function of time, has a particular type of shape, with its exact shape being determined by one or more parameters which are estimated using the observed data. Some commonly used distributions in survival analysis are the *exponential* (Figure 4, left panel, black line), the *Weibull* (Figure 4, left panel) and the *log-logistic* distribution (Figure 4, right panel).

The exponential distribution is the simplest with a single parameter to be estimated and a special case of the Weibull distribution. The choice of parametric family to be used depends on the shape of the distribution. Fitting an exponential distribution to a set of data assumes that the underlying hazard function is constant in time, that is, it assumes that the occurrence of events in time is totally random. A Weibull distribution allows a monotonic (either continuously increasing or decreasing hazard) and a log-logistic distribution allows either a monotonic or a unimodal hazard function.

When using parametric models we make assumptions the plausibility of which should be investigated. For example if the event of interest is death of any cause and the time origin is an individual's birth (that is, our time scale is age), then using an exponential model is not a valid option, as the instantaneous all-cause death probability (i.e. the hazard) is unlikely to be constant with age.

Such a model can be fitted to a set of survival data in order to summarize the features of the data. It may also facilitate the comparison of two or more sets of data. Parametric models can be used in regression analyses of survival data when the effects of other variables on survival are to be investigated. Estimation of the parameters can be done using maximum likelihood. The parameter estimates are found by differentiation the log likelihood with respect to the unknown parameters, setting the derivatives to zero and solving the resulting equations with respect to the parameters.

Regression models

One of the objectives of the analysis of survival data might be to examine whether survival times are related to other features. Regression models can be used to assess the effect of covariates on the outcome. These are similar to regression analyses for other types of outcomes, such as linear regression for a continuous numeric outcome or logistic regression for binary outcomes.

Two common types of regression models for survival data, classified by the way in which covariates are assumed to affect the survival times are the *Cox proportional hazards model* and the *accelerated failure time* (or *accelerated life*) *model*.

Cox proportional hazards model

A Cox proportional hazards model[6] has the form

$$h(t;x) = h_0(t)e^{\beta x}$$

where $h_0(t)$ is the *baseline hazard*, x is a covariate and β is a parameter to be estimated, representing the effect of the covariate on the outcome. The baseline hazard is the hazard when, in the case of a single covariate, the covariate is equal to zero. The main assumption implied is the *proportional hazards assumption*, which is that the *hazard ratio*, that is the ratio of the hazard function to the baseline hazard, is constant over time. The use of the exponential function ensures that the hazard is positive.

The quantity that is estimated from a Cox proportional hazards model is interpreted as *relative*, rather than *absolute*, *risk*. The covariates are assumed to have an additive effect on the log hazard ratio (the natural logarithm of the hazard ratio). The interpretation of the parameter β is that for each unit increase in the covariate x, the hazard is multiplied by e^{β}. In the special case of x taking the values 0 or 1 to represent to groups, say A and B, group B has e^{β} times the risk of group A.

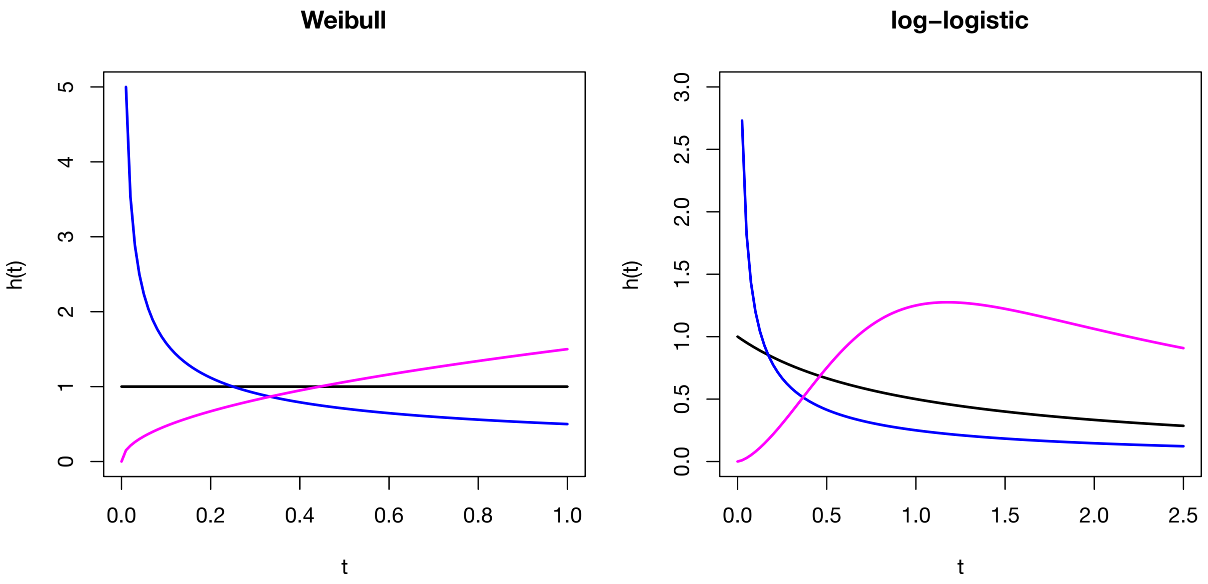

Figure 4 Hazard function of Weibull (left) and log-logistic (right) distributions. The different colours represent different parameter values yielding different shapes of the hazard function within each family of distributions.

As in all regression models, more than one variables can be included in a Cox proportional hazards model, to adjust for the effects of other variables. The multivariable version of the Cox proportional hazards model can be written as

$$h(t;x_1,\ldots,x_p) = h_0(t)e^{\beta_1 x_1+\ldots+\beta_p x_p}.$$

In this case the baseline hazard is the hazard for an individual with all his/her covariates equal to zero ($x_1 = \ldots = x_p = 0$). The effect of variable x_k is interpreted as follows: for each unit increase in x_k and all other covariates held fixed, the hazard is multiplied by e^{β_k}.

Estimation

The unknown parameters β in a Cox proportional hazards model can be estimated using the *partial likelihood*,

$$PL(\beta) = \prod_{t_i:\text{ event at } t_i} \frac{h_0(t)e^{\beta x_{(t_i)}}}{\sum\limits_{j\in R_{(t_i)}} h_0(t)e^{\beta x_j}},$$

where the product is taken over ordered event times, $R_{(t_i)}$ is the *risk set* at time t_i, that is, the subjects which are still in the sample just prior to time t_i, and $x_{(t_i)}$ is the value of x for the subject which had an event at time t_i. The probabilities that the partial likelihood consists of are the probabilities that the individual who has the event at a given event time is actually the individual who had it out of all individuals at risk at that time. This implicitly assumes that there is only one event occurring at each event time, that is, that there are no tied event times. There are methods for dealing with tied event times which are implemented by statistical software packages, incorporated into the standard implementations of Cox models. The baseline hazards cancel out from the numerator and denominator and do not need to be estimated and therefore the partial likelihood is simplified to

$$PL(\beta) = \prod_{t_i:\text{ event at } t_i} \frac{e^{\beta x_{(t_i)}}}{\sum\limits_{j:\ t_j\geq t_i} e^{\beta x_j}}.$$

This is why the Cox proportional hazards model is referred to as a *semi-parametric* method, that is, a method in which survival times are assumed to be related to the explanatory variables in a particular way, but no assumptions are made on the overall shape of the survival times, that is the shape of the hazard function need not be specified.

The partial likelihood can be treated as a likelihood. Standard errors for the estimate of β are based on asymptotic results.[7]

Example. In the previous example, let x represent the number of lymph nodes with detectable cancer. Suppose that we fit the model $h(t;x) = h_0(t)e^{\beta x}$. We find that the estimate of β is $\widehat{\beta} = 0.092$ and the standard error of $\widehat{\beta}$ is $\text{se}(\widehat{\beta}) = 0.0088$. Thus the hazard ratio is $\widehat{e^{\beta}} = 1.10$. Therefore for each additional lymph node with cancer, the risk is multiplied by 1.10. A 95% confidence interval for the hazard ratio is (1.08, 1.12), derived using a normal approximation for the log hazard ratio. The p-value is close to zero, which provides strong evidence that the number of lymph nodes with detectable cancer is associated with death from colon cancer.

The plausibility of the proportional hazards assumption should be checked. It can be checked graphically or by including a time-dependent effect and examining its significance. A quick graphical check is to plot the scaled *Schoenfeld residuals*, implemented by most statistical packages. These show whether the effect β as a function of time varies with time. If it varies substantially with time, it suggests that the proportional hazards assumption might not be plausible. There is a χ^2 test derived from Schoenfeld residuals which formally tests departures from the null hypothesis of proportionality, thus with a significant result being evidence of non-proportionality. Another graphical way of assessing the plausibility of the proportional hazards assumption for simple cases with a single explanatory variable is to plot $\log\{-\log\ S(t;x)\}$ against time for different values of the explanatory variable. It can be easily shown by substituting the form of the survival function that under the proportional hazards assumption the curves should be separated by a constant vertical deviation (that is, they should be parallel), equal to the effect β of the explanatory variable. Thus seeing the separation between the curves substantially vary by time and even more seeing the curves cross suggests that the assumption is not appropriate. For more than one explanatory variables the plot could be done on combinations of possible values of the variables.

A possible solution to a model for which the proportional hazards assumption seems not to be plausible is to change the set of covariates included in the model or alternatively to *stratify* by a categorical variable. *Stratification* in this context means to partition individuals into strata and to allow a different baseline hazard $h_{0k}(t)$ in each stratum k but to still assume that the effect of the covariates on the outcome is the same for the entire dataset. It might also be used if it is thought that there are differences between the groups defined by the strata which cannot be fully accounted for by the covariates. Thus any differences are absorbed in the stratum-specific baseline hazards and a single effect is estimated for each covariate. A stratified model does not allow comparisons to be made between strata. Another more complex option when the proportional hazards assumption is not met is to include a *time-dependent effect*, which is an extension of the standard Cox model, or fit separate models for different parts of the time scale. Cox models with time-dependent effects are beyond the scope of this article. In some cases it may be more appropriate to use a different type of model.

More specialized ways of assessing some aspects of model fit include the *Cox–Snell residuals*, *martingale residuals* and *deviance residuals*.

Parametric proportional hazard models

The Weibull model (and thus the exponential, being a special case of the Weibull) can be used in regression as a proportional hazards model. If we assume that the explanatory variable acts multiplicatively on the hazard, starting from the survival function of a Weibull model and replacing the baseline hazard with the hazard that includes the explanatory variable effect, we end up with a survival function which has the form of a Weibull distribution but with different parameters. Therefore the Weibull model belongs to the family of proportional hazards models. It can be shown that the Weibull model can also be written as an accelerated life model.

Accelerated life models

An accelerated life model (or accelerated failure time model) is a regression model in which the survival function is assumed to have the same shape for all individuals and explanatory variables are assumed to affect survival by changing the speed with which individuals move on the curve. That is, some individuals move across it more slowly or more quickly than others. So instead of having the hazard multiplied by a quantity as in a proportional hazards model, here the survival time is multiplied by a quantity. This can be written as $T = \psi T_0$ or $T = T_0 e^{-\beta}$. This yields the survival function

$$S(t;x) = S_0\left(e^{\beta x} t\right),$$

where $S_0(\cdot)$ is the 'baseline' survival of an individual with his/her explanatory variable taking the value zero. The factor e^{β} is now called the *acceleration factor*, that is, it represents how faster or slower an individual would move on the survival curve for a unit increase in the explanatory variable x. If the acceleration factor is greater than 1 then individuals with higher values of x will tend to have earlier event times, whereas if it is less than 1 then individuals with higher values of x will tend to have later event times, that is their survival times will be longer.

The hazard and density function can be deduced from the survival function. Unlike proportional hazards models, accelerated life models are usually fully parametric. The log-logistic model is an example of an accelerated life model.

In some applications an accelerated life model may have a more directly explicable interpretation. For example in modelling event time outcomes related to ageing, one might prefer to deduce effects to be interpreted as accelerating or decelerating the underlying ageing process rather than proportional increases in risk throughout time. Accelerated life models can also be used to yield estimates of changes in life expectancy.

Choice of time scale

In some examples the time scale might be possible to be specified in more than one ways. For example, in a study in which individuals enter the study as adults and are followed up until the development of a particular disease, it is possible to use age as the time scale, that is having an individual's birth as the time origin and disease onset as the endpoint, allowing for delayed entry such that individuals are not yet 'at risk' until they enter the study. Alternatively, the time origin might be chosen to be the time of entry into the study. The choice is usually based on subject-matter considerations, the better option being the one under which individuals are as similar as possible with respect to their underlying risk of event at the time origin. This is discussed by Ref. 8.

Time-dependent explanatory variables

So far we have assumed that an explanatory variable is measured once and represents either a feature of an individual at a fixed time point with respect to the time origin, or a feature which remains unchanged for the time of observation, for example a patient's sex or occupation. However, the value of an explanatory variable might be changing with time. We call such a variable a *time-dependent* (alternatively *time-varying* or *time-updated*) *explanatory variable*. The Cox proportional hazards model can accommodate such variables. In practice this would require splitting time into discrete time units, for example years, and assigning a value for each variable at each such time unit during which an individual is at risk. Therefore time-to-event data with time-dependent explanatory variables are represented by multiple observations per individual, each representing a time unit during which the individual was at risk.

A time-dependent explanatory variable is different from a time-dependent effect mentioned previously. A Cox proportional hazards model with a time-dependent explanatory variable can be written as

$$h(t;x(t)) = h_0(t)e^{\beta x(t)}.$$

This gives an effect on the hazard at time t of the explanatory variable at *that* time t. It is possible to have different formulations, such as allowing a time-lagged variable. The effect of a variable that is increased by one unit for each time unit in the data, such as age of entry into a study, is the same as that of the non-updated version of the variable, given that its effect on the log hazard ratio is linear. For variables that change over time but are only measured a number of times throughout the follow-up time, it is common to keep their most recent value as the 'current' value. However care is needed if whether a variable is measured at a given time point may depend on factors which may also influence its value and the outcome.

Some other points

Other regression models

A different, not as commonly used, type of regression model in which effects are assumed to act additively on the hazard is *Aalen's additive hazard model*.[9] The interpretation of the effects that this model yields is more complicated that the other types of models discussed above. Muirhead and Darby[10] compared proportional and additive hazard models. There also exists a hybrid multiplicative—additive model, called the *Cox—Aalen model*.[11]

Competing risks

Another issue that may arise in a study with time-to-event data is the problem of *competing risks*. Competing risks are different types of failure or event which may occur to the individuals being studied. For example, in an epidemiological prospective cohort study where the effect of a risk factor on death from cancer is being studied, some participants may die from cardiovascular disease. Care is needed in handling such data, as in some cases censoring at the time of competing events may invalidate the assumption of non-informative censoring. There may be bias which is likely to be greater when the hazard of the competing event is greater.

Suppose there are K competing events, assumed to be *absorbing*, that is, once an individual has one event they cannot have another one. Then for each type of event a *cause-specific* *hazard function* can be defined. The *cumulative incidence function* (or *subdistribution function*) is the probability of having an event of type k ($k = 1, \ldots, K$) before time t. The *subdistribution hazard*[12] is the instantaneous rate of occurrence of event type k among individuals who are either event-free or have had an event other than type k. This approach does not incorporate time-dependent explanatory variables and it does not allow testing of whether explanatory variables have the same effect on different event types. Another concept is the *cause-specific Kaplan–Meier curve*, which is the probability of survival from cause k beyond time t if all competing failure types have been eliminated but type k has remained unchanged. This is usually unrealistic and difficult to interpret. A Cox proportional hazards model on cause-specific hazards can be defined. The interpretation of effects of variables estimated from such a model are effects on the hazard among individuals who are eligible to have an event of type k. For a discussion of these issues as well as *multi-state models*, which are modelling transitions between many states and extend time-to-event analyses to event-history analyses, see Ref. 13.

Frailty models

Frailty models are used for incorporating heterogeneity between individuals using a *random effect*. See for example Ref. 14.

Recurrent events

Frailty models can also be used for modelling data with *recurrent events*, such as one or more hospitalizations of the same individual for a particular relapsing condition. Other methods are also available. See for example Ref. 15.

Risk prediction

Another potential objective of the analysis of survival data is *absolute risk prediction*. That is, given an individual's explanatory features we may want to predict the probability of an event occurring to that individual, either as a function of time or within a given time period. In some cases we may have time-updated explanatory variables and may want to update predictions after 'baseline' (i.e. study entry, not to be confused with baseline meaning 'reference group'). This is sometimes called *dynamic prediction*. Approaches for dynamic prediction include joint modelling of the time-varying (or longitudinal) covariates and survival, multi-state modelling and *landmarking*, which is a method for making predictions for survival conditionally on things occurring after baseline. For methods for dynamic prediction see Ref. 16.

Special sampling schemes/study designs

So far the discussion was mainly focused on the analysis of prospective cohort data. Data from different study designs or non random sampling techniques may require some modification to the methods presented here. One such special sampling technique is *matching*, where individuals are matched according to one or more features mainly in order to reduce variability and avoid confounding due to those features. Individuals are commonly matched in pairs. For methods for analysis of paired survival data see for example Ref. 17.

Another situation is that sometimes we have a prospective cohort study but one or more explanatory variables cannot be measured on the full set of individuals, usually due to cost or other practical considerations. Then it might be possible for them to be measured in a subset of the full cohort. One way of selecting such a sample is to identify a set of *cases*, that is, individuals who had the event of interest, and for each identify one or more *controls*, that is individuals who have not yet had the event at the event time of their case. This is called a *nested case-control study*. Another way of selecting such a sample is to identify a set of cases and then select a random sample of the cohort, called the *subcohort*, from the data at study entry, ignoring any information collected after baseline. This is called a *case-subcohort study* (or sometimes a *case-cohort study*). Nested case-control and case-subcohort studies are analysed using simple extensions to the basic methods for time-to-event data. For methods for design and analysis of such studies see Ref. 18.

Some books on survival analysis are Refs. 19–24. ◆

REFERENCES

1 Moertel CG, Fleming TR, MacDonald JS, et al. Fluorouracil plus Levamisole as an effective adjuvant therapy after resection of stage III colon carcinoma: a final report. *Ann Intern Med* 1995; **122:** 321–6.

2 R Core Team. R: a language and environment for statistical computing. 2015. Vienna: R Foundation for Statistical Computing, http://www.R-project.org/.

3 Therneau T. *A package for survival analysis in S. Version 2.38* 2015. URL http://CRAN.R-project.org/package=survival.

4 Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958; **58:** 457–81.

5 Greenwood M. The natural duration of cancer. Reports on public health and medical subjects**33**. London: Her Majesty's Stationery Office, 1926; 1–26.

6 Cox DR. Regression models and life-tables. *J R Stat Soc B* 1972; **34:** 187–220.

7 Cox DR. Partial likelihood. *Biometrika* 1975; **62:** 269–76.

8 Farewell VT, Cox DR. A note on multiple time scales in life testing. *J R Stat Soc C* 1979; **28:** 73–5.

9 Aalen OO. A linear regression model for the analysis of life times. *Stat Med* 1989; **8:** 907–25.

10 Muirhead C, Darby S. Modelling the relative and absolute risks of radiation-induced cancers. *J R Stat Soc A* 1987; **150:** 83–118.

11 Scheike TH, Zhang MJ. An additive–multiplicative Cox–Aalen regression model. *Scand J Stat* 2002; **29:** 75–88.

12 Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc* 1999; **94:** 496–509.

13 Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med* 2007; **26:** 2389–430.

14 Hougaard P. Frailty models for survival data. *Lifetime Data Anal* 1995; **1:** 255–73.

15 Rogers JK, Pocock SJ, McMurray JJ, et al. Analysing recurrent hospitalizations in heart failure: a review of statistical methodology, with application to CHARM-Preserved. *Eur J Heart Fail* 2014; **16:** 33–40.

16 van Houwelingen H, Putter H. Dynamic prediction in clinical survival analysis. CRC Press, 2011.

17 Kartsonaki C, Cox DR. Some matched comparisons of two distributions of survival time. *Biometrika* 2016; **103:** 219–24.

18 Keogh RH, Cox DR. Case-control studies. Cambridge: Cambridge University Press, 2014.

19 Cox DR, Oakes DO. Analysis of survival data. London: Chapman and Hall, 1984.

20 Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. 2nd edn. New York: John Wiley & Sons, 2002.

21 Collett D. Modelling survival data in medical research. 2nd edn. Boca Raton: Chapman & Hall/CRC, 2003.

22 Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York: Springer, 2000.

23 Kleinbaum DG, Klein M. Survival analysis: a self-learning text. 2nd edn. New York, NY: Springer, 2005.

24 Moeschberger ML, Klein JP. Survival analysis: techniques for censored and truncated data. 2nd edn. New York: Springer Science and Business Media, Inc, 2003.