ST3242: Introduction to Survival Analysis

Semester 2, 2010–2011

Alex R Cook

You are free to copy, distribute and display these notes, and to make derivative works, under the following conditions: (i) you give the original author (i.e. A. R. Cook) credit; and (ii) if you alter, transform, or build upon this work, you may distribute the resulting work only under a licence identical to this one.
Course details and syllabus

Contact details

Dr Alex Cook

Drop-in office hours: Monday 09.00–11.00, level 6 room 113 (or by appointment).

Email: alex.richard.cook at gmail.com


IVLE: There is also an IVLE web page, though the important information will appear on the webpage above.

Please do ask questions!

Suggested reading

Two good books are Kleinbaum and Klein (2005) and Hosmer et al. (2008). Both books cover the material of this course, and examples will be drawn from them (and other sources). Kleinbaum & Klein is much easier to read and better laid out; Hosmer et al. covers more mathematical details, but is quite heavy going. Both are expensive (about $150!), and although the NUS
Co-op declined my request to stock Kleinbaum and Klein (2005), they will be stocking Hosmer et al. (2008) and will order Kleinbaum and Klein (2005) if you request it. The library has books on survival analysis that cover the course material.

Course organisation

Lectures will be held on Mondays 12.00–13.30 and Thursdays between 12.00–13.30pm at LT34, unless tutorial allocation screws up the timetable or the room transpires to be too large.

In addition there will be tutorials held weekly, with details to be confirmed later once the class size has been confirmed.

Assessment

Mid-term test: 15%

Projects: 25%

Final exam: 60%

Syllabus

1. Introduction to introduction to survival analysis. You will learn to understand censoring, one of the key differences between statistical inference for survival data and other kinds of data, and the properties of the survival and hazard functions. We will briefly cover likelihood for parametric models: more parametric modelling will follow later in the course. Topics in this chapter are:
(a) censoring & truncation;
(b) hazard & survival functions; and
(c) likelihood for parametric models.

2. **Kaplan–Meier estimates of the survival function.** You will learn the derivation and use of the Kaplan–Meier (KM) non-parametric estimate of the survival function. You will learn how to plot the KM estimator by hand and on the computer, and to test for differences in the survival function between subpopulations. You will learn how to

(a) estimate KM by hand;
(b) calculate $V\{S(t)\}$ and CIs;
(c) estimate mean survival time;
(d) do KM on computer; and
(e) compare KM curves.

3. **Cox’s proportional hazards model.** You will learn the motivation behind Cox’s semi-parametric proportional hazards model for understanding the effects of covariates on survival, know how to fit the model on the computer, to test for significant effects and adjust the KM estimate of survival accordingly. You will learn to

(a) understand the Cox PHM;
(b) estimate coefficients from first principles on computer;
(c) test equality of coefficients using the likelihood-ratio (LR) test;
(d) fit model and do LR test on a computer;
(e) estimate adjusted $S(t)$; and
(f) find CIs for hazard ratio.

4. **Modelling with Cox’s PHM.** You will learn the art of modelling survival, by choosing what terms to put in a model and assessing relative and absolute goodness-of-fit. This chapter covers:

(a) guidelines for choosing covariates in models;
(b) assessing PH assumption using:
   - graphical methods and
• goodness-of-fit tests; and
(c) extending the model to include time-varying terms.

5. Extensions of Cox’s PHM. You will learn two significant extensions to Cox’s PHM, using stratification to account for covariates that would otherwise invalidate the proportional hazards assumption and to extend the model to incorporate time-varying covariates or covariates that have a time-varying effect on survival. We will also cover the application of the Cox model to data with unusual censoring. The chapter covers:

(a) stratified Cox;
(b) extended Cox model; and
(c) Cox PHM with unusual censoring/truncation.

6. Parametric survival analysis. We will recap the basics of parametric models for survival data and cover common survival distributions. Two popular types of parametric analyses will be considered: accelerated failure time models and frailty models. The material to be covered will be:

(a) common distributions;
(b) accelerated failure time models;
(c) frailty models.
Chapter 1

Introduction to introduction to survival analysis

Survival analysis is the study of the time it takes for events to happen. It is frequently used in medical statistics—to understand the time to remission of cancer, to the next psychiatric episode, to death after release from hospital—although it has many other applications, in ecology, business, and recidivism, for example. Survival analytical techniques provide a way to understand via analogies of regression the effect of predictors on the development of the event of interest. Time to event data are complicated by the censoring that often results from patient drop-out or termination of studies, and this makes analysis of such data markedly different to other experimental or observational data.

These notes will guide you in learning to distinguish the different kinds of censored data, to fit and compare non-parametric summaries of survival data via the Kaplan-Meier method, to develop semi-parametric regression models using the Cox proportional hazards model and extensions thereof, to assess the goodness of fit of such regression models, and to fit and evaluate parametric alternatives. By the end of the course, you should be able to develop appropriate models of real survival data and to present their results to a professional standard.
CHAPTER 1. INTRODUCTION

This chapter introduces the key factors that distinguish survival data from other sorts of data commonly encountered in statistics—censoring, which often results from patients dropping out of long studies; and the survival and hazard functions, which are used to summarise survival data—before briefly introducing parametric models for survival data. After this chapter, we will look at non- and semi-parametric methods to understand survival data, returning to parametric models only in the final chapter of the course.

1.1 Introduction to censoring and truncation

1.1.1 Definitions

Censoring is when an observation is incomplete due to some random cause. The cause of the censoring must be independent of the event of interest if we are to use standard methods of analysis.

Truncation is a variant of censoring which occurs when the incomplete nature of the observation is due to a systematic selection process inherent to the study design.

Although censoring and truncation do arise in non-survival data, for instance, when test readings cannot be made below a certain minimum threshold, censoring in particular is a key issue in survival analysis, affecting almost all analysis.

1.1.2 Examples of censored data

Imagine a study in which lung cancer patients are recruited to test the effect of a drug on their survival from lung cancer, e.g. by randomly allocation consenting patients in a hospital to receive the new drug or the current standard of care equiprobably.
• Patient A takes part in the study until her death at age $T_A$. Her survival time is *uncensored*.

• Patient B takes part in the study until age $T_B$. He then leaves the study. His survival time is *censored*: we know it is *at least* $T_B$ but we don’t know it precisely.

• Patient C takes part in the study until age $T_C$. She then is hit by a car and dies. Her survival time *with regard to the event of interest*, namely death through lung cancer, is also *censored*: we know it is *at least* $T_C$ but the unfortunate accident has censored it.

### 1.1.3 Types of censoring

• The commonest form of censoring is *right censoring*. Both examples in the previous section were of right censoring. Here, the subject is followed until some time, at which the event has yet to occur, but then takes no further part in the study. This may be because:

  – the subject dies from another cause, independently of the cause of interest;
  – the study ends while the subject survives; or
  – the subject is lost to the study, by dropping out, moving to a different area, etc.

If our data contain only uncensored and right-censored data, we can represent all individuals by the triple ($i, t_i, \delta_i$):

  – $i$ indexes subjects,
  – $t_i$ is the time at which the death or censoring event occurs to individual $i$, and
  – $\delta_i$ is an indicator: $\delta_i = 1$ if $i$ is uncensored and $\delta_i = 0$ if censored.

Most methods in the course apply to right censored data only.

• *Left censoring* is much rarer. This occurs when the event of interest has already occurred at the observation time, but it is not known exactly when. Examples of left censoring include:
• Infection with a sexually-transmitted disease such as HIV/AIDS;
• Onset of a pre-symptomatic illness such as cancer; and
• Time at which teenagers begin to drink alcohol, later forgotten.

• Interval censoring is both extremely common and very rare. This occurs when the exact time the event occurs is not known precisely, but an interval bounding this time is known. If the interval is very short, e.g. 1d, it is common to ignore this form of censoring and pick one end point of the interval consistently. Examples of interval censoring include:
  • Infection with a sexually-transmitted disease such as HIV/AIDS with regular testing (e.g. annually); and
  • Failure of a machine during the Chinese New Year.

• Left truncation occurs when, due to the structure of the study design, we can only observe those individuals whose event time is greater than some truncation threshold. This threshold may be the same for all individuals or may be random. For example, in actuarial life tables, it is common to ignore those who die in the womb. As a further example: imagine you wish to study how long people who have been hospitalised for a heart attack survive taking some treatment at home. The start time is taken to be the time of the heart attack. Only those individuals who survive their stay in hospital are able to be included in the study.

• Right truncation is when only individuals with event times less than some threshold are included in the study. If you ask a group of smoking school pupils at what age they started smoking, you necessarily have truncated data, as individuals who start smoking after leaving school are not included in the study. Similarly, if a drug developed in 1990 is compared with a placebo to analyse the survival times of (dead) cancer patients, all times to death must be \( \leq 20y \).

1.1.4 Quiz: what kind of censoring or truncation?

Try to identify which of the following scenarios is associated with each kind of censoring or truncation.
• Leukæmia patients are given a drug or a placebo. Survival time is the duration from remission to relapse. The study ends at 52 weeks with some patients yet to relapse.

• The age at which children are able to count from 1–10 at school. Some children are already able to count before joining school.

• Death times at a retirement community.

• Early in the AIDS outbreak, patients with AIDS were recruited to study the time from infection with HIV to development of AIDS. At the time of the study, many people were infected with HIV but had not yet developed symptoms of AIDS.

• NUS students are asked the age at which they first tried marijuana.
  - Some answer never.
  
  - Some answer with an exact age.
  
  - Some report using it but forget when.
1.1.5 Importance of independence

It is vital that the cause of any censoring or truncation be independent of the failure event of interest. If not, the methodology in the course cannot be used. Here is an example of a violation of the independence assumption: suppose patients are recruited who have been operated on for lung cancer. Survival is defined to be the time after diagnosis until death from lung cancer. Some patients may die of other cancers before dying of lung cancer. This is not independent of death due to lung cancer since cancers often spread to other parts of the body, so those dying of other cancers are more likely to have died soon after due to lung cancer than others in the sample.

If independence is in doubt, you may be able to recast the problem in some way. In the example above, survival could be redefined as the time of death from any cancer. Defining the event of interest as death from any cause is clear and unambiguous, although in quantifying effect sizes it must be remembered that the effect of the treatment or association with the predictor is on overall mortality, not lung cancer mortality, and findings reported as such. Another possibility is to do a sensitivity analysis: imagining two extreme scenarios in which deaths due to other cancers (say) are set to be (i) independent censoring events, or (ii) pretended to be deaths due to lung cancer, and then seeing whether your results are consistent.

1.1.6 Fitting parametric models with censoring

Although censored and truncated data are only partially observed, we do not throw them away. Excluding them:

- increases the variance in our estimates by reducing the sample size; and
- introduces bias to the results.
CHAPTER 1. INTRODUCTION

Fortunately, for parametric models, it is easy to construct the likelihood function in the presence of censoring.

- If $T$ is distributed according to some known family of distributions with unknown parameters $\theta$, then its density and distribution functions are $f(t|\theta)$ and $F(t|\theta) = \int_0^t f(\tau|\theta) d\tau$, respectively.

- Let $\mathcal{R}$ be the set of right-censored, $\mathcal{L}$ of left-censored, $\mathcal{I}$ of interval-censored and $\mathcal{U}$ of uncensored observations.

- If we observe $D = (\mathcal{R}, \mathcal{L}, \mathcal{I}, \mathcal{U}, T)$ where $T = \{t_i : i = 1, \ldots, n\}$, the log-likelihood function is

$$
\log L(D|\theta) = \sum_{i \in \mathcal{U}} \log \frac{f(t_i)}{w_i} + \sum_{i \in \mathcal{R}} \log \frac{1 - F(t_i)}{w_i} + \sum_{i \in \mathcal{L}} \log \frac{F(t_i^L)}{w_i} + \sum_{i \in \mathcal{I}} \log \frac{F(t_i^R) - F(t_i^L)}{w_i}
$$

where $t_i^L$ and $t_i^R$ are the lower and upper bounds on the infection time for interval censored data resp. The weights $w_i$ are 1 in the absence of truncation. If individual $i$ is truncated within the interval $(Y_i^L, Y_i^R)$, it changes to $w_i = F(Y_i^R) - F(Y_i^L)$. If $i$ is left-truncated only, $F(Y_i^R) = F(\infty) = 1$; if right-truncated only, $F(Y_i^L) = F(0) = 0$.

1.2 The survival and hazard functions

1.2.1 Introduction

In most statistical analyses, interest is in the probability density (or mass) function and the cumulative distribution function (CDF) of the random variable of interest, and of various derivations thereof, such as the moments. Although these are important, in survival analysis more emphasis is placed upon the survival and hazard functions.

The survival function is the probability of surviving to some time $t$. The hazard function is the instantaneous rate at which failure occurs at time $t$. 
1.2.2 Survival function

The survival function, $S(t)$, is very similar to the CDF, but is the probability of surviving at least to time $t$, rather than of failing by time $t$. A very common way to estimate it non-parametrically was given by Kaplan and Meier (1958), which allows its estimation even in the presence of censoring. Being estimable in the presence of censoring is perhaps the main reason $S(t)$ is favoured to the density $f(t)$.

**Definition**

If $T$ is the time of failure, then $S(t) = p(T > t)$ is the survival function. It is defined on the domain $t \in [0, \infty)$; as a probability, it has range $S(t) \in [0, 1]$.

**Properties**

- $S(0) = 1$
- $\lim_{t \to \infty} S(t) = 0$
- $S(t_a) \geq S(t_b) \iff t_a \leq t_b$
- $S(t) = 1 - F(t) = \int_t^\infty f(\tau) \, d\tau$

Typically, the population survival function is smooth, though estimates of it are not.

**Estimating the survival function**

We would like to estimate the survival function underlying observed data. There are two main approaches one can take.


- **Parametric**
  If you have a fully specified parametric model for $T$ with parameters $\theta$, then the estimated $S(t)$ follows from the estimate, $\hat{\theta}$. Parameters can be estimated using standard methods, such as *maximum likelihood* (should know already), from the *posterior distribution* of $\theta$ given the data within the Bayesian paradigm (see ST4234 Bayesian Statistics!), the method of moments, etc. An example using maximum likelihood will follow.

- **Non-parametric**
  If you cannot justify a parametric model, you may use non-parametric methods, such as the popular Kaplan–Meier estimate, to estimate $S(t)$. This will be covered in the next lecture.

**Example**

If $T \sim \text{We}(\kappa, \lambda)$ with $f(t) = \lambda \kappa t^{\kappa-1} \exp(-\lambda t^\kappa)$ then $S(t) = \ldots$

---

### 1.2.3 Hazard function

The *hazard function*, $h(t)$, is the *instantaneous failure rate*. Its formal definition is:
CHAPTER 1. INTRODUCTION

Definition

\[ h(t) = \lim_{\Delta t \to 0} \frac{p(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \]

Can we write it as a function of \( S(t) \)?

What about \( S(t) \) as a function of \( h(t) \)?
Note that if $h(t) = h \forall t \geq 0$, then $S(t) = \exp\{-ht\}$. Recall that if $X \sim \text{Exp}(\lambda)$ then $p(X \leq x) = 1 - \exp(-\lambda x)$. Thus, a constant hazard is equivalent to exponentially distributed lifetimes, that is, lifetimes that are memoryless. This means that if two individuals have survived until ages $t_a$ and $t_b$, the remainder of their life time has the same distribution, even if $t_a \ll t_b$. It is for this reason that temporally homogeneous (i.e. constant) hazards are unrealistic for most populations, although they do come up in stochastic processes (see ST4238 Stochastic Processes 2!). **Task:** Show that $p(T > t_a + t|T > t_a) = p(T > t_b + t|T > t_b)$ if $T \sim \text{Exp}(h)$ (see tutorial 1).

### Hazard properties

- $h(t) \geq 0 \forall t \geq 0$
- $\int_0^{\infty} h(t) \, dt = \infty$ otherwise there is a non-zero probability that some individuals will survive for ever.

To understand the second point, imagine $\int_0^{\infty} h(t) \, dt = H < \infty$. Then $S(\infty) = e^{-H} > 0$. Note that $h(t)$ is not a probability, and so, like PDFs, $h(t) > 1$ is perfectly acceptable.

#### 1.2.4 Cumulative hazard function

The cumulative hazard function is $H(t) = \int_0^t h(\tau) \, d\tau$. We shan’t cover this.
1.2.5 Behaviour of hazard function

**Constant** $h(t)$

appropriate for time until next case of influenza in a non-seasonal country?

**Increasing** $h(t)$

appropriate for cancer patients that are not responding to treatment, machines with more faults with age, death rates among adult animals.
**Decreasing** $h(t)$

![Graphs showing decreasing functions](image)

Appropriate for survival following surgery.

**Rising and falling** $h(t)$

![Graphs showing rising and falling functions](image)

Appropriate for survival following tuberculosis infection.

**Falling and rising** $h(t)$

![Graphs showing falling and rising functions](image)
Appropriate for lifespan of animals, machines where faults surface immediately or with age.

### 1.2.6 Parametric models

The key theoretical constraint is that the domain of $f(t)$ must be $\mathbb{R}^+$. This rules out the Normal distribution, for instance. Suitable distribution families include:

**Exponential**

\[
\begin{align*}
    f(t) &= \\
    F(t) &= \\
    S(t) &= \\
    h(t) &= \lambda
\end{align*}
\]

**Gompertz**

\[
\begin{align*}
    f(t) &= \\
    F(t) &= \\
    S(t) &= \\
    h(t) &= \lambda \kappa^t
\end{align*}
\]

**Weibull**

\[
\begin{align*}
    f(t) &= \\
    F(t) &= \\
    S(t) &= \\
    h(t) &= \lambda \kappa t^{\kappa-1}
\end{align*}
\]
Log-logistic

\[
f(t) = \frac{\lambda \kappa t^{\kappa - 1}}{1 + \lambda t^\kappa}
\]

\[
F(t) = 1 - \frac{1}{1 + \lambda t^\kappa}
\]

\[
S(t) = \frac{1}{1 + \lambda t^\kappa}
\]

\[
h(t) = \frac{\lambda \kappa t^{\kappa - 1}}{1 + \lambda t^\kappa}
\]

1.2.7 Example
It is generally true that the density function for $t$ is the product of the hazard and survival functions, as:

$$f(t) = dF(t) = -dS(t) = -dS(t) = h(t)S(t).$$

For non-censored observations, the likelihood contribution from individual $i$ is $f(t_i) = h(t_i)S(t_i)$ which can be interpreted as the probability that no event occurs until time $t_i$ (i.e. $S(t_i)$) times the instantaneous rate of failure at time $t_i$ ($h(t_i)$). The total likelihood function for the parameters $\theta$ of the model for $T$ is thus $f(t|\theta) = \prod_{i=1}^{n} f(t_i)$. This can be maximised via calculus or numerically to find the MLE of $\theta$, or can be fed into a Bayesian routine to obtain the posterior distribution of $\theta$.

Example: $h(t) = ae^{bt}$ and no censoring

$$\log f(t|a, b) = \sum_{i=1}^{n} \log f(t_i)$$

$$= \sum_{i=1}^{n} \log h(t_i) + \log S(t_i)$$

$$= \sum_{i=1}^{n} \log a + bt_i + \frac{a}{b}(1 - e^{bt_i}).$$

The derivatives are not very nice and so we would maximise numerically, e.g. by simulated annealing or a grid search.

If there were censoring, we’d take the same approach (calculus or numerical) but on the censored likelihood function given last time.

Confidence intervals for $\theta$ can be obtained in the usual way, i.e. by appealing to the asymptotic distribution of the MLE: $f(\hat{\theta}) \rightarrow N(\theta, I(\hat{\theta})^{-1})$ as $n \rightarrow \infty$ where

$$I(\hat{\theta}) = E\left\{-\frac{\partial^2 \log L(\hat{\theta})}{\partial \theta \partial \theta^T}\right\}$$
is the Fisher information.

1.2.8 Comparison of parametric and non-parametric

The advantage of assuming a parametric form for the hazard, survival or density functions is that \( h(t) \) and \( S(t) \) are both fully specified in terms of \( \theta \), smooth estimates of \( h \) and \( S \) can be found that draw strength from other times, and tests for differences in parameters are more powerful. The disadvantage is that it is very easy to be too sure of your model, to draw inappropriate conclusions by distorting the signal from some times to others. The alternative for descriptive methods and hypothesis testing is to use non-parametric methods, which are very popular in survival analysis. These are covered in the next chapter. Semi-parametric methods make some (testable) assumptions and allow modelling of the effect of covariates and are covered thereafter. We return to parametric methods later in the course.