An Introduction to Survival Analysis[*]

Mark Stevenson

EpiCentre, IVABS, Massey University

December 2007

Contents

---

[*]Notes for MVS course 195.721 Analysis and Interpretation of Animal Health Data. `http://epicentre.massey.ac.nz`

1 General principles

Survival analysis is the name for a collection of statistical techniques used to describe and quantify time to event data. In survival analysis we use the term 'failure' to define the occurrence of the event of interest (even though the event may actually be a 'success' such as recovery from therapy). The term 'survival time' specifies the length of time taken for failure to occur. Situations where survival analyses have been used in epidemiology include:

- Survival of patients after surgery.

- The length of time taken for cows to conceive after calving.

- The time taken for a farm to experience its first case of an exotic disease.

1.1 Describing time to event

Death density

When the variable under consideration is the length of time taken for an event to occur (e.g. death) a frequency histogram can be constructed to show the count of events as a function of time. A curve fitted to this histogram produces a death density function $f(t)$, as shown in Figure 1. If we set the area under the curve of the death density function to equal 1 then for any given time t the area under the curve to the left of t represents the proportion of individuals in the population who have experienced the event of interest. The proportion of individuals who have died as a function of t is known as the cumulative death distribution function and is called $F(t)$.

Survival

Considering again the death density function shown in Figure 1. The area under the curve to the right of time t is the proportion of individuals in the population who have survived to time t, $S(t)$. $S(t)$ can be plotted as a function of time to produce a survival curve, as shown in Figure 2. At $t = 0$ there have been no failures so $S(t) = 1$. By day 15 all members of the population have failed and $S(t) = 0$. Because we use counts of individuals present at discrete time points, survival curves are usually presented in step format.

Hazard

The instantaneous rate at which a randomly-selected individual known to be alive at time (t - 1) will die at time t is called the conditional failure rate or instantaneous hazard, $h(t)$. Mathematically, instantaneous hazard equals the number that fail between time t and time $t + \Delta(t)$ divided by the size of the population at risk at time t, divided by $\Delta(t)$. This gives the proportion of the population present at time t that fail per unit time.

Instantaneous hazard is also known as the force of mortality, the instantaneous death rate, or the failure rate. An example of an instantaneous hazard curve is shown in Figure 3. Figure 3 shows the the weekly probability of foot-and-mouth disease occurring in two farm types in Cumbria (Great Britain) in 2001. You should interpret this curve in exactly the same way you would an epidemic curve. The advantage of this approach is that it shows how disease risk changes over time, correcting for changes in the size of the population that might occur over time (something that should be considered when dealing with foot-and-mouth disease data).
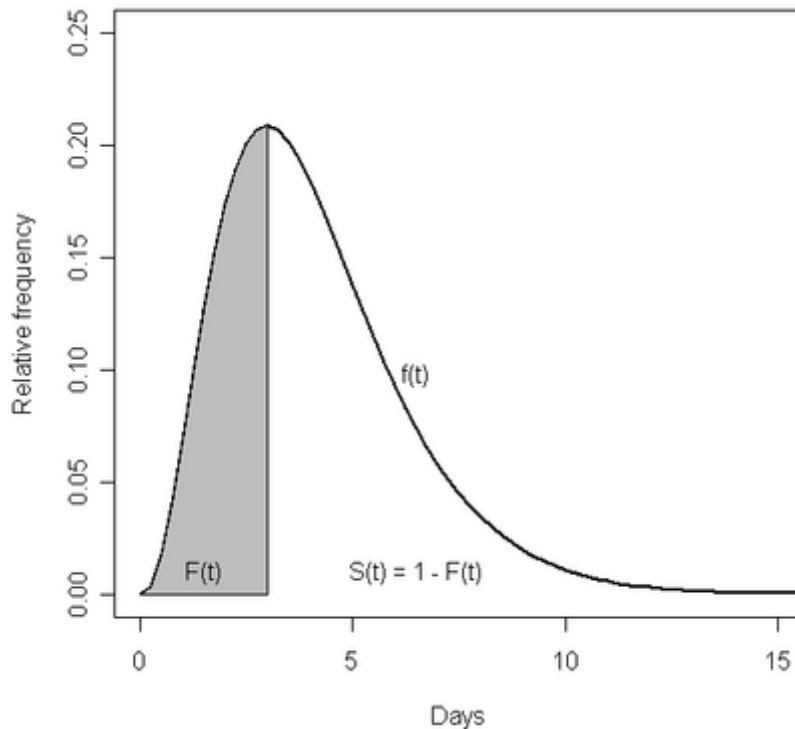
Figure 1: Line plot $f(t)$ (death density) as a function of time. The cumulative proportion of the population that has died up to time t equals $F(t)$. The proportion of the population that has survived to time t is $S(t) = F(t) - 1$.

The cumulative hazard (also known as the integrated hazard) at time t, $H(t)$ equals the area under the hazard curve up until time t. A cumulative hazard curve shows the (cumulative) probability that the event of interest has occurred up to any point in time.

1.2 Censoring

In longitudinal studies exact survival time is only known for those individuals who show the event of interest during the follow-up period. For others (those who are disease free at the end of the observation period or those that were lost) all we can say is that they did not show the event of interest during the follow-up period. These individuals are called censored observations. An attractive feature of survival analysis is that we are able to include the data contributed by censored observations right up until they are removed from the risk set. The following terms are used in relation to censoring:

- Right censoring: a subject is right censored if it is known that failure occurs some time *after* the recorded follow-up period.

- Left censoring: a subject is left censored it it is known that the failure occurs some time *before* the recorded follow-up period. For example, you conduct a study investigating factors influencing days to first oestrus in dairy cattle. You start observing your population (for argument's sake) at 40 days after calving but find that several cows in the group have already had an oestrus event. These cows are said to be left censored at day 40.

- Interval censoring: a subject is interval censored if it is known that the event occurs between two times, but the exact time of failure is not known. In effect we say 'I know that the event occurred between date A and date B: I know that the event occurred, but I don't know exactly when.' In an
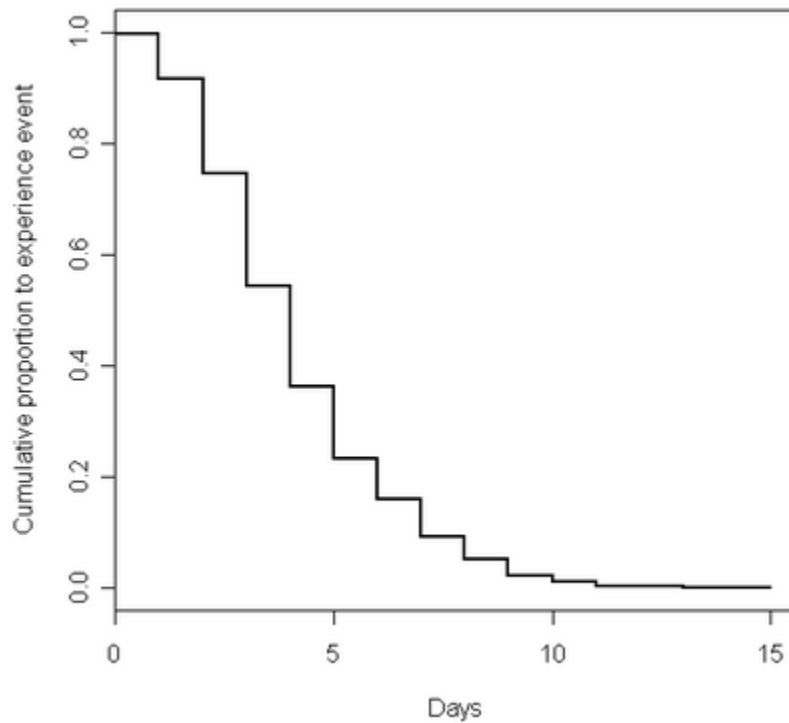
4

Figure 2: Survival curve showing the cumulative proportion of the population who have 'survived' (not experienced the event of interest) as a function of time.

observational study of EBL seroconversion you sample a population of cows every six months. Cows that are negative on the first test and positive at the next are said to have seroconverted. These individuals are said to be interval censored with the first sampling date being the lower interval and the second sampling date the upper interval.

We should distinguish between the terms censoring and truncation (even though the two events are handled the same way analytically). A truncation period means that the outcome of interest *cannot possibly* occur. A censoring period means that the outcome of interest *may* have occurred. There are two types of trunction:

- Left truncation: a subject is left truncated if it enters the population at risk some stage after the start of the follow-up period. For example, in a study investigating the date of first BSE diagnosis on a group of farms, those farms that are established after the start of the study are said to be left truncated (the implication here is that there is no way the farm can experience the event of interest before the truncation date).

- Right truncation: a subject is right truncated if it leaves the population at risk some stage after the study start (and we know that there is no way the event of interest could have occurred after this date). For example, in a study investigating the date of first foot-and-mouth disease diagnosis on a group of farms, those farms that are pre-emptively culled as a result of control measures are right truncated on the date of culling.

Consider a study illustrated in Figure 4. Subjects enter at various stages throughout the study period. An 'X' indicates that the subject has experienced the outcome of interest; a 'O' indicates censoring. Subject A experiences the event of interest on day 7. Subject B does not experience the event during the
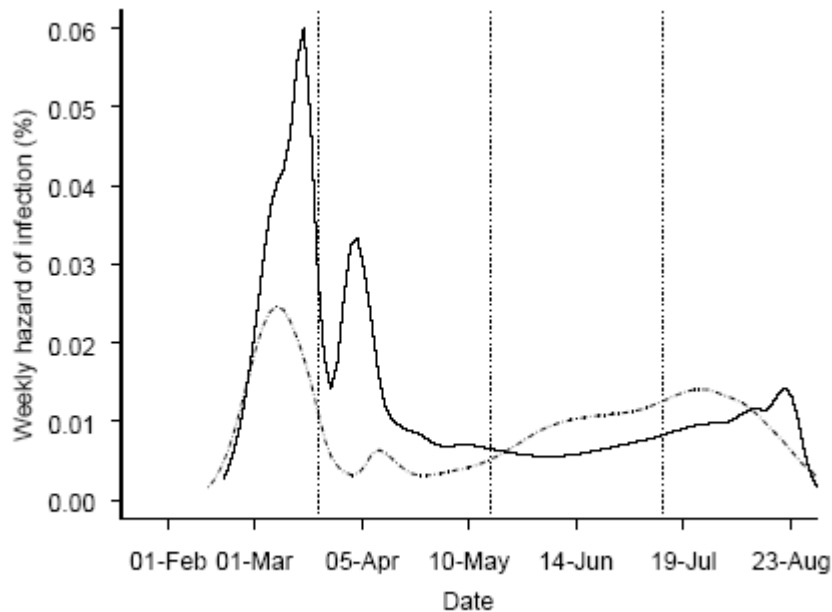
Figure 3: Weekly hazard of foot-and-mouth disease infection for cattle holdings (solid lines) and 'other' holdings (dashed lines) in Cumbria (Great Britain) in 2001. Reproduced from Wilesmith et al. (2003).

study period and is right censored on day 12 (this implies that subject B experienced the event sometime after day 12). Subject C does not experience the event of interest during its period of observation and is censored on day 10. Subject D is interval censored: this subject is observed intermittantly and experiences the event of interest sometime between days 5 – 6 and 7 – 8. Subject E is left censored — it has been found to have already experienced the event of interest when it enters the study on day 1. Subject F is interval truncated: there is no way possible that the event of interest could occur to this individual between days 4 – 6. Subject G is left truncated: there is no way possible that the event of interest could have occurred before the subject enters the study on day 3.

2 Non-parametric survival

Once we have collected time to event data, our first task is to describe it — usually this is done graphically using a survival curve. Visualisation allows us to appreciate temporal pattern in the data. It also helps us to identify an appropriate distributional form for the data. If the data are consistent with a parametric distribution, then parameters can be derived to efficiently describe the survival pattern and statistical inference can be based on the chosen distribution. Non-parametric methods are used when no theoretical distribution adequately fits the data. In epidemiology non-parametric (or semi-parametric) methods are used more frequently than parametric methods.

There are three non-parametric methods for describing time to event data: (1) the Kaplan-Meier method, (2) the life table method, and (3) the Nelson-Aalen method.
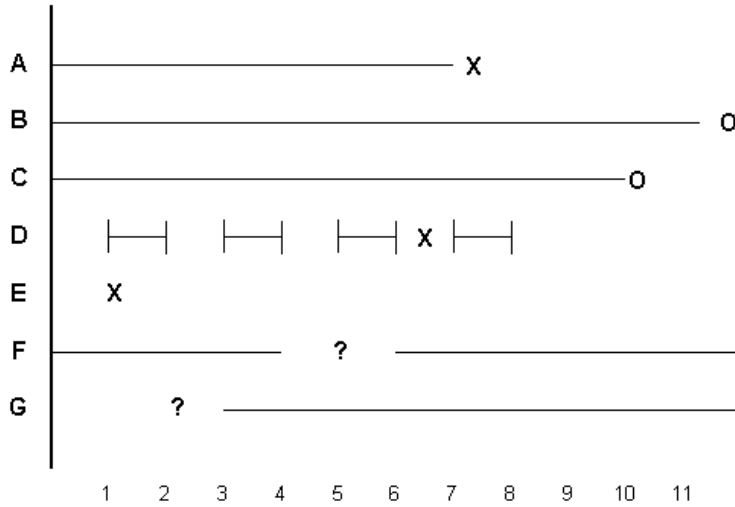
Figure 4: Left-, right-censoring, and truncation (Dohoo, Martin and Stryhn 2003).

2.1 Kaplan-Meier method

The Kaplan-Meier method is based on individual survival times and assumes that censoring is independent of survival time (that is, the reason an observation is censored is unrelated to the cause of failure). The Kaplan-Meier estimator of survival at time t is shown in Equation 1. Here t_j, $j = 1, 2, ..., n$ is the total set of failure times recorded (with t^+ the maximum failure time), d_j is the number of failures at time t_j, and r_j is the number of individuals at risk at time t_j. A worked example is provided in Table 1. Note that: (1) for each time period the number of individuals present at the start of the period is adjusted according to the number of individuals censored and the number of individuals who experienced the event of interest in the previous time period, and (2) for ties between failures and censored observations, the failures are assumed to occur first.

$$\hat{S}(t) = \prod_{j:t_j \leq t} \frac{(r_j - d_j)}{r_j}, \text{ for } 0 \leq t \leq t^+ \tag{1}$$

Table 1: Details for calculating Kaplan-Meier survival estimates as a function of time.

Time	Start	Fail	Censored	At risk	Surv prob	Cumulative survival
	n_j	d_j	w_j	r_i	$P_j = (r_j - d_j)/r_j$	$S_j = P_j \times P_{j-1}$
0	31	2	3	31 - 3 = 28	(28 - 2) / 28 = 0.93	0.93 × 1.00 = 0.93
1	26	1	2	26 - 2 = 24	(24 - 1) / 24 = 0.96	0.96 × 0.93 = 0.89
2	23	1	2	23 - 2 = 21	(21 - 1) / 21 = 0.95	0.95 × 0.89 = 0.85
3	20	1	2	20 - 2 = 18	(18 - 1) / 18 = 0.94	0.94 × 0.85 = 0.80
etc						

Table 2: Details for calculating life table survival estimates as a function of time.

Time	Start	Fail	Censored
	n_i	d_i	w_i
0 to 1	31	3	4
2 to 3	24	2	4
etc			

Time	Failure prob	Survival prob	Cumulative survival
	$q_i = d_i/[n_i - (w_i/2)]$	$p_i = 1 - q_i$	$S_i = p_i \times S_{i-1}$
0 to 1	3 / [31 - (4/2)] = 0.10	1 - 0.10 = 0.90	0.90 × 1 = 0.90
2 to 3	2 / [24 - (4/2)] = 0.09	1 - 0.09 = 0.91	0.90 × 0.91 = 0.82
etc			

2.2   Life table method

The life table method (also known as the actuarial or Cutler Ederer method) is an approximation of the Kaplan-Meier method. It is based on grouped survival times and is suitable for large data sets. Calculation details are shown in Table 2.

The life table method assumes that subjects are withdrawn randomly throughout each interval — therefore, on average they are withdrawn half way through the interval. This is not an important issue when the time intervals are short, but bias may introduced when time intervals are long. This method also assumes that the rate of failure within an interval is the same for all subjects and is independent of the probability of survival at other time periods. Life tables are produced from large scale population surveys (e.g. death registers) and are less-frequently used these days (the Kaplan-Meier method being preferred because it is less prone to bias).

2.3   Flemington-Harrington estimator

Instantanous hazard is defined as the proportion of the population present at time t that fail per unit time. The cumulative hazard at time t, $H(t)$ is the summed hazard for all time up to time t. The relationship between cumulative hazard and survival is as follows:

$$H(t) = -ln[S(t)], \text{ or } S(t) = e^{-H(t)} \tag{2}$$

The Nelson-Aalen estimator of cumulative hazard at time t is defined as:

$$\hat{H}(t) = \sum_{j:t_j \leq t} \frac{d_j}{r_j}, \text{ for } 0 \leq t \leq t^+ \tag{3}$$

The Flemington-Harrington estimate of survival can be calculated using the Nelson-Aalen estimate of cumulative hazard using the relationship between survival and cumulative hazard described in Equation 2.

2.4   Worked examples

An Australian study by Caplehorn and Bell (1991) compared retention in two methadone treatment clinics for heroin addicts. A patient's survival time was determined as the time in days until the patient dropped out of the clinic or was censored at the end of the study. The two clinics differed according to their overall treatment policies. Interest lies in identifying factors that influence retention time: clinic, maximum daily methadone dose, and presence of a prison record.

Kaplan-Meier method

Figure 5 is a Kaplan-Meier survival curve showing the cumulative proportion of addicts retained in the clinics over time. Figure 5 shows that the rate of loss of patients over time is relatively constant and that approximately 15% remain in treatment by 1000 days post admission.
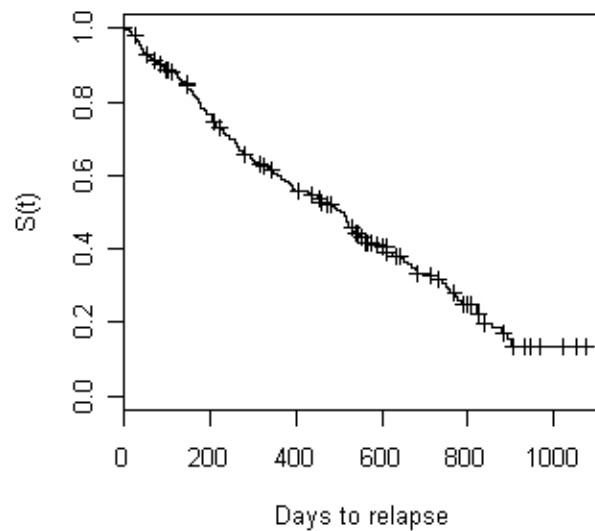

Figure 5: Kaplan-Meier survival curve showing the cumulative proportion of heroin addicts retained in two methadone treatment clinics (Caplehorn and Bell 1991).

Figure 5 was produced using the following code:

```
library(survival)
dat <- read.table("addict.csv", header = TRUE, sep = ",")
```

Plot Kaplan-Meier survival function of days from discharge from clinic to relapse:

```
addict.km <- survfit(Surv(stop, status) ~ 1, conf.type = "none", type =
"kaplan-meier", data = dat)
plot(addict.km, xlab = "Days to relapse", ylab = "S(t)")
```

Kaplan-Meier survival function with confidence intervals:

```
addict.km <- survfit(Surv(stop, status) ~ 1, type = "kaplan-meier", data =
dat)
plot(addict.km, xlab = "Days to relapse", ylab = "S(t)", conf.int = TRUE)
```

Kaplan-Meier survival function of days to relapse, stratifying by clinic:

```
addict.km <- survfit(Surv(stop, status) ~ clinic, type = "kaplan-meier", data
= dat)
plot(addict.km, xlab = "Days to relapse", ylab = "S(t)", lty = c(1,2),
legend.text = c("Clinic 1","Clinic 2"), legend.pos = 0, legend.bty = "n")
```

Kaplan-Meier survival function of days to relapse, stratifying by methadone dose:

```
dat$newdose[dat$dose < 60] <- 0
dat$newdose[dat$dose >= 60] <- 1
addict.km <- survfit(Surv(stop, status) ~ newdose, type = "kaplan-meier", data
= dat)
plot(addict.km, xlab = "Days to relapse", ylab = "S(t)", lty = c(1,2),
legend.text = c("Low dose methadone","High dose methadone"), legend.pos = 0,
legend.bty = "n")
```

Flemington-Harrington estimator

```
addict.km <- survfit(Surv(stop, status) ~ 1, type = "kaplan-meier", data =
dat)
addict.fh <- survfit(Surv(stop, status) ~ 1, type = "fleming-harrington", data
= dat)

par(pty = "s", mfrow = c(1,2))
plot(addict.km, xlab = "Days to relapse", ylab = "S(t)", conf.int = FALSE)
plot(addict.fh, xlab = "Days to relapse", ylab = "S(t)", conf.int = FALSE)
```

With this data set the difference between the Kaplan-Meier and the Flemington-Harrington estimate of survival is not obvious. A closer comparison of the two functions:

```
tmp <- as.data.frame(cbind(km = addict.km$surv, fh = addict.fh$surv))
head(tmp)
tail(tmp)
```

Instantaneous hazard

```
addict.km <- survfit(Surv(stop, status) ~ 1, conf.type = "none", type =
"kaplan-meier", data = dat)
```

Work out the proportion that fail at each evaluated time period:

```
prop.fail <- addict.km$n.event/addict.km$n.risk
```

Work out the length of time over which these failure occur:

```
time <- addict.km$time
time
[1] 2, 7, 13, 17, ..., 1076

time0 <- c(0, time[-length(time)])
time0
[1] 0, 2, 7, 13, ..., 1052
```

Divide `prop.fail` by the time interval over which those failures occur (that is, `time - time0`) to get the probability of failing per unit time, i.e. the instantaneous hazard:

```
haz <- prop.fail/(time - time0)
```

Plot the result:

```
plot(time, haz, ylim = c(0,0.06), type = "s", xlab = "Days to relapse", ylab =
"h(t)")
lines(lowess(time[-1], haz[-1], f = 0.10))
```

Tidier plot:

```
plot(time, haz, type = "n", xlab = "Days to relapse", ylab = "h(t)", ylim =
c(0,0.06))
lines(lowess(time[-1], haz[-1], f = 0.10))
```

A simpler way to do it. You will need to have the epicentre package installed:

```
library(epicentre)
addict.haz <- epi.insthaz(addict.km)
plot(addict.haz$time, addict.haz$est, xlab = "Days to relapse", ylab = "h(t)",
, ylim = c(0,0.06), type = "s")
lines(lowess(addict.haz$time, addict.haz$est, f = 0.10))
```

Cumulative hazard

```
addict.km <- survfit(Surv(stop, status) ~ 1, conf.type = "none", type =
"kaplan-meier", data = dat)
plot(addict.km, fun = "cumhaz", xlab = "Days to relapse", ylab = "H(t)", lty =
c(1,2))
```

Compare instantaneous hazard with cumulative hazard:

```
par(pty = "s", mfrow = c(1,2))
plot(addict.km, fun = "cumhaz", xlab = "Days to relapse", ylab = "H(t)")
plot(addict.haz$time, addict.haz$est, xlab = "Days to relapse", ylab = "h(t)",
ylim = c(0,0.06), type = "s")
```

3 Parametric survival

On some occassions the pattern of survivorship for our study subjects follows a predictable pattern. In this situation, parametric distributions can be used to describe time to event. An advantage of using a parametric distribution is that we can reliably predict time to event well after the period during which

events occurred for our observed data. Several parametric distributions are used to describe time to event data. Each parametric distribution is defined by a different hazard function, as shown in Table 3.

Table 3: Parametric survival distributions used in epidemiology.

Distribution	$f(t)$[a]	$h(t)$[b]	$H(t)$[c]	$S(t)$[d]
Exponential	$\lambda \exp[-\lambda t]$	λ	λt	$\exp[-\lambda t]$
Weibull	$\lambda p t^{p-1} \exp[-\lambda t^p]$	$\lambda p t^{p-1}$	λt^p	$\exp[-(\lambda t)^p]$
Gompertz	$a \exp[bt] \exp[-a/b (\exp[bt] - 1)]$	$a \exp[bt]$	$a/b (\exp[bt] - 1)$	$\exp[-a/b (\exp[bt] - 1)]$
Log-logistic	$abt^{b-1} / (1 + at^b)^2$	$(abt^{b-1}) / (1 + at^b)$	$\log(1 + at^b)$	$(1 + at^b)^{-1}$

[a] $f(t)$ probability density.
[b] $h(t)$ instantaneous hazard.
[c] $H(t)$ cumulative hazard.
[d] $S(t)$ survival.

The Gompertz distribution provides a convenient way of describing survival in human subjects and is frequently used in demography. The Gompertz distribution can be generalised to the Gompertz-Makeham distribution by adding a constant to the instantaneous hazard: $h(t) = c + a \exp(bt)$.

As a general approach to the analysis of time to event data we should plot the hazard function for the observed data and determine whether or not it is consistent with a parametric distribution. If the data follows a parametric distribution, parametric methods are preferred to non-parametric methods for describing and quantifying factors that influence time to event. In veterinary epidemiology, the most important parametric forms are the exponential and Weibull distributions.

3.1 The exponential distribution

The exponential distribution is described by the mean, λ. A feature of the exponential distribution is that the instantaneous hazard does not vary over time (Figure 6). Observed survival distributions can be checked for consistency with the exponential distribution by plotting instantaneous hazard as a function of time: exponential distributions in this case will yield a straight line. Alternatively, the log of cumulative hazard can be plotted as a function of the log of time: exponential distributions will yield a 45° line.
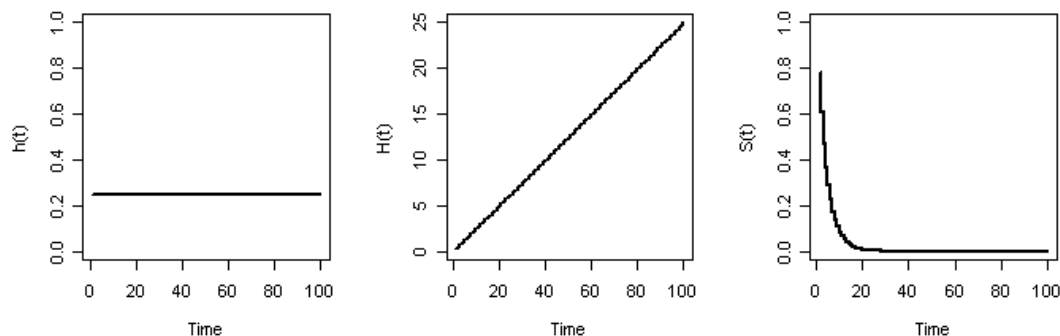

Figure 6: Instantaneous hazard, cumulative hazard and survival as a function of time for the exponential distribution. In this example $\lambda = 0.25$.

3.2 The Weibull distribution

The Weibull distribution is described by a scale parameter λ and shape parameter p. If p is less than 1 instantaneous hazard monotonically decreases with time, if p equals 1 instantaneous hazard is constant over time (equivalent to the exponential distribution) and if p is greater than 1 instantaneous hazard increases with time. Figure 7 is an example of a Weibull distributed survival pattern with $p < 1$. Time to event data can be checked for consistency with the Weibull distribution by plotting the log cumulative hazard as a function of log time: Weibull distributions in this case will yield a straight line.
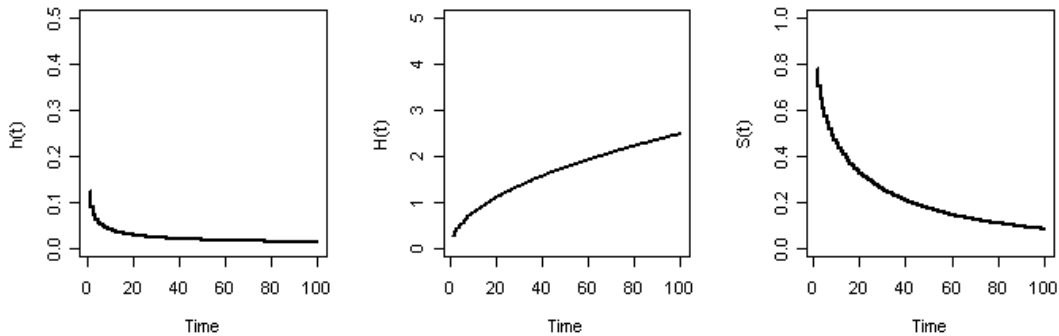
Figure 7: Instantaneous hazard, cumulative hazard and survival as a function of time for the Weibull distribution. In this example $\lambda = 0.25$ and $p = 0.5$.

3.3 Worked examples

The exponential distribution

Figure 6 was produced using the following code:

```
t <- seq(from = 1, to = 100, by = 1)
lambda = 0.25
ht <- lambda
Ht <- lambda * t
St <- exp(-lambda * t)

par(mfrow = c(2,3), pty = "s")
plot(t, rep(ht, times = length(t)), ylim = c(0, 1), lwd = 2, type = "s", xlab
= "Time", ylab = "h(t)")
plot(t, Ht, ylim = c(0, 25), lwd = 2, type = "s", xlab = "Time", ylab =
"H(t)")
plot(t, St, ylim = c(0, 1), lwd = 2, type = "s", xlab = "Time", ylab = "S(t)")
```

The Weibull distribution

Figure 7 was produced using the following code:

```
t <- seq(from = 1, to = 100, by = 1)
lambda = 0.25; p = 0.5
ht <- lambda * p * t^(p - 1)
Ht <- lambda * t^p
St <- exp(-lambda * t^p)

par(mfrow = c(2,3), pty = "s")
plot(t, ht, ylim = c(0, 0.5), lwd = 2, type = "s", xlab = "Time", ylab =
"h(t)")
plot(t, Ht, ylim = c(0, 5), lwd = 2, type = "s", xlab = "Time", ylab = "H(t)")
plot(t, St, ylim = c(0, 1), lwd = 2, type = "s", xlab = "Time", ylab = "S(t)")
```

Plots of hazard using different values of `lambda` and p:

```
t <- seq(from = 0, to = 10, by = 0.1)
lambda <- 1; p05 <- 0.5; p10 <- 1.0; p15 <- 1.5; p30 <- 3.0
h05 <- lambda * p05 * (lambda * t)^(p05 - 1)
h10 <- lambda * p10 * (lambda * t)^(p10 - 1)
h15 <- lambda * p15 * (lambda * t)^(p15 - 1)
h30 <- lambda * p30 * (lambda * t)^(p30 - 1)

plot(t, h05, type = "l", ylim = c(0, 6), xlab = "Time", ylab = "h(t)", lty =
1, lwd = 2)
lines(t, h10, lty = 2, lwd = 2)
lines(t, h15, lty = 3, lwd = 2)
lines(t, h30, lty = 4, lwd = 2)
legend(4, 6, legend = c("lambda = 1, p = 0.5", "lambda = 1, p = 1.0", "lambda
= 1, p = 1.5", "lambda = 1, p = 3.0"), lty = c(1,2,3,4), lwd = c(2,2,2,2), bty
= "n", cex = 0.75)
```

Comparison of Kaplan-Meier and Weibull estimates of survival:

```
setwd("D:\\TEMP")
library(survival)
dat <- read.table("addict.csv", header = TRUE, sep = ",")
```

Fit parametric (Weibull) and non-parametric (Kaplan-Meier) survival functions to the observed data:

```
addict.we <- survreg(Surv(stop, status) ~ 1, dist = "weib", data = dat)
addict.km <- survfit(Surv(stop, status) ~ 1, conf.type = "none", type =
"kaplan-meier", data = dat)
```

Using the Weibull distribution μ (the intercept) = $-\log(\lambda)$ and σ (scale) = 1 / p. Thus the scale parameter $\lambda = \exp(-\mu)$ and p = 1 / σ. See Venables and Ripley p 360 and Tableman and Kim p 78 for further details.

```
p <- 1 / addict.we$scale
lambda <- exp(-addict.we$coeff[1])
t <- 1:1000
St <- exp(-(lambda * t)^p)
addict.we <- as.data.frame(cbind(t = t, St = St))
```

Compare the two estimates of survival:

```
plot(addict.km, xlab = "Days to relapse", ylab = "Cumulative proportion to
experience event")
lines(addict.we$t, addict.we$St, lty = 2)
legend(x = "topright", legend = c("Kaplan-Meier", "Weibull"), lty = c(1,2),
bty = "n")
```

The Weibull distribution provides an adequate fit to the observed data up to day 500, then appears to underestimate survivorship. Cumulative hazard plots can provide an alternative method for assessing the appropriateness of a parametric approach to describe survivorship. Here we plot cumulative hazard as a function of time (to check for consistency with the exponential distribution) and log cumulative hazard as a function of log time (to check for consistency with the Weibull distribution). Both plots show departure from linearity.

```
addict.km <- survfit(Surv(stop, status) ~ 1, conf.type = "none", type =
"kaplan-meier", data = dat)
Ht <- -log(addict.km$surv)
t <- addict.km$time

par(pty = "s", mfrow = c(1,2))
plot(t, Ht, type = "s", xlim = c(0, 1000),xlab = "Days to relapse", ylab =
"Cumulative hazard", lwd = 2)
plot(log(t), log(Ht), type = "s", xlim = c(0, 7), xlab = "Days to relapse
(log)", ylab = "Cumulative hazard (log)", lwd = 2)
```

4 Comparing survival distributions

It is frequently of interest to compare the survival of one group of study subjects with another.

- Did animals survive longer in one herd compared with another?
- Did disease take longer to develop in one region of a country compared with another?
- Did patients survive longer after one therapy compared with another?

In addition to providing useful information about how time to event distributions differ among groups, separate survival curves for different levels of covariates provide an effective screening process that helps one to identify factors that are influential in determining survival. Once influential factors are screened using these methods their influence can then be tested using multivariate analyses.

When there are no censored observations, standard non-parametric tests can be used to compare two survival distributions. If the groups are independent, a Wilcoxon or Mann-Whitney U test may be used. If the groups are dependent the Sign Test may be used.

4.1 The log-rank test

The log-rank test (also known as the Mantel log-rank test, the Cox Mantel log-rank test, and the Mantel-Haenszel test) is the most commonly used test for comparing survival distributions. It is applicable to data where there is progressive censoring and gives equal weight to early and late failures. It assumes that hazard functions for the two groups are parallel. The test takes each time point when a failure event occurs and a 2×2 table showing the number of deaths and the total number of subjects under follow up is created. For each table the observed deaths in each group, the expected deaths and the variance of the expected number are calculated. These quantities are summed over all tables to yield a χ^2 statistic with 1 degree of freedom (known as the Mantel-Haenszel or log-rank test statistic). The log-rank test calculations also produce for each group the observed to expected ratio which relates the number of deaths observed during the follow up with the expected number under the null hypothesis that the survival curve for that group would be the same as that for the combined data.

4.2 Other tests

Breslow's test (also known as Gehan's generalised Wilcoxon test) is applicable to data where there is progressive censoring. It is more powerful than the log-rank test when the hazard functions are not parallel and where there is little censoring. It has low power when censoring is high. It gives more weight to early failures.

The Cox Mantel test is similar to the log-rank test. It is applicable to data where there is progressive censoring. More powerful than Gehan's generalised Wilcoxon test. The Peto and Peto modification of the Gehan-Wilcoxon test is similar to Breslow's test and is used where the hazard ratio between groups is not constant. Cox's F test is more powerful than Breslow's test if sample sizes are small.

4.3 Worked examples

```
library(survival)
setwd("D:\\TEMP")
dat <- read.table("addict.csv", header = TRUE, sep = ",")

Kaplan-Meier survival function of days to relapse, stratifying by clinic:

addict.km <- survfit(Surv(stop, status) ~ clinic, type = "kaplan-meier", data
= dat)
plot(addict.km, xlab = "Days to relapse", ylab = "Cumulative proportion to
experience event", lty = c(1,2), legend.text = c("Clinic 1", "Clinic 2"),
legend.pos = 1, legend.bty = "n")

In the survdiff function the argument rho = 0 returns the log-rank or Mantel-Haenszel test, rho = 1 returns the Peto and
Peto modification of the Gehan-Wilcoxon test. Mantel-Haenszel test:

survdiff(Surv(stop, status) ~ clinic, data = dat, na.action = na.omit, rho =
0)

Peto and Peto modification of the Gehan-Wilcoxon test:

survdiff(Surv(stop, status) ~ clinic, data = dat, na.action = na.omit, rho =
1)
```

5 Non-parametric and semi-parametric models

Survival models are used to quantify the effect of one or more explanatory variables on failure time. This involves specification of a linear-like model for the log hazard. A parametric model based on the

exponential distribution may be parameterised as follows:

$$\log h_i(t) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} \tag{4}$$

or, equivalently:

$$h_i(t) = \exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}) \tag{5}$$

In this case the constant α represents the log-baseline hazard since $\log h_i(t) = \alpha$ when all the x's are zero. The Cox proportional hazards model is a semi-parametric model where the baseline hazard $\alpha(t)$ is allowed to vary with time:

$$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} \tag{6}$$

$$h_i(t) = h_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}) \tag{7}$$

If all of the x's are zero the second part of the above equation equals 1 so $h_i(t) = h_0(t)$. For this reason the term $h_0(t)$ is called the baseline hazard function. With the Cox proportional hazards model the outcome is described in terms of the hazard ratio. We talk about the hazard of the event of interest at one level of an explanatory variable being a number of times more (or less) than the hazard of the specified reference level of the explanatory variable.

Assumptions of the Cox proportional hazards model are as follows:

- The ratio of the hazard function for two individuals with different sets of covariates does not depend on time.

- Time is measured on a continuous scale.

- Censoring occurs randomly.

Table 4 presents the results of a Cox proportional hazards regression model for the Caplehorn `addict` data set (Caplehorn and Bell 1991). Here the authors have quantified the effect of clinic, methadone dose, and prison status on the daily hazard of relapse (re-using heroin). Clinic is a categorical variable with Clinic 1 as the reference category. The results of the model show that, compared with patients from Clinic 1 and after adjusting for the effect of methadone dose and prison status, Clinic 2 patients had 0.36 (95% CI 0.24 – 0.55) times the daily hazard of relapse. Similarly, for unit increases in the daily dose of methadone, after adjusting for the effect of clinic and the presence of a prison record there was a 4% decrease in daily hazard of relapse (95% CI 3% – 5%).

5.1   Model building

Selection of covariates

We now discuss how a set of variables are selected for inclusion in a regression model of survival. Begin with a thorough univariate analysis of the association between survival time and all important covariates. For categorical variables this should include Kaplan-Meier estimates of the group-specific survivorship functions. Tabulate point and interval estimates of the median and quartiles of survival time. Use one or more of the significance tests to compare survivorship among the groups defined by the variable under investigation. Continuous covariates should be broken into quartiles (or other biologically meaningful groups) and the same methods applied to these groups.

Table 4: Cox proportional hazards regression model showing the effect of clinic, methadone dose and prison status on the daily hazard of relapse (adapted from Caplehorn and Bell 1991).

Variable	Subjects	Failed	Coefficient (SE)	P	Hazard (95% CI)
Clinic				< 0.01 [a]	
Clinic 1	163	122			1.0
Clinic 2	75	28	-1.0091 (0.2147)		0.36 (0.24 − 0.55) [b]
Prison record				0.06	
Absent	127	81			1.0
Present	111	69	0.3146 (0.1672)		1.37 (0.98 − 1.90)
Dose	238	150	-0.0352 (0.0064)	< 0.01	0.96 (0.95 − 0.98)

[a] The significance of inclusion of the two clinic variables in the model.
[b] Intepretation: compared with the reference category (patients from Clinic 1) after adjusting for the effect of methadone dose and prison status patients from Clinic 2 had 0.36 (95% CI 0.24 − 0.55) times the daily hazard of relapse.
SE: standard error.
CI: confidence interval.

Tied events

A central assumption in survival analysis is that time is continuous. Sometimes (particularly in veterinary epidemiological research) the outcome of interest is not measured on a continuous scale and outcome events may occur simultaneously (e.g. service number when conception occurred). When the number of tied events is large, approximate methods yield regression coefficients that are biased towards zero. There are three common methods for dealing with ties:

1. Breslow approximation. There is a contribution to the partial likelihood from each of the tied failure times. For each failure time, the risk set comprises all subjects failing at or after the failure time. This includes all subjects whose failure times are tied with that of the subject contributing to the numerator.

2. Efron approximation. In the Breslow approximation, if m subjects share the same survival time, they all contribute to the risk set for each of the m failure times as if each one of the m subjects failed, all others were still alive. In the Efron approximation, the contribution to the denominator from the m subjects with tied survival times is weighted down by a factor of $(m - k)/m$ for the kth term.

3. Exact partial likelihood. Assuming that no two subjects ever failed simultaneously (this would be the case if we measured the time of failure down to milliseconds), there is a true (unknown) unique ordering of the tied survival times. The exact partial likelihood can be obtained by taking the sum (or average) of the partial likelihoods for all possible orderings of the tied survival times. Computationally intensive.

Fitting a multivariable model

A multivariable model should contain at the outset all covariates significant in the univariate analyses at the P = 0.20 to 0.25 level and any others that are thought to be of clinical importance. You should also include any covariate that has the potential to be an important confounder.

Following the fit of the multivariable model, use the P values from the Wald tests of the individual coefficients to identify covariates that might be deleted from the model. The partial likelihood ratio test

should confirm that the deleted covariate is not significant. Also check if removal of a covariate produces a 'significant' change (say 20%) in the coefficients of the covariates remaining in the model. Continue until no covariates can be deleted from the model. At this point, work backwards and add each of the deleted covariates back into the model one at a time — checking that none of them are significant or show evidence of being a confounder.

Check the scale of continuous covariates

The next thing is to examine the scale of the continuous covariates in the preliminary model. Here we need to check that the covariate is linear in its log hazard. Replace the continuous covariate with three design variables using Q1, Q2, and Q3 as cutpoints. Plot the estimated coefficients for the design variables versus the midpoint of the group. A fourth point is included at zero using the midpoint of the first group. If the correct scale is linear, then the line connecting the four points should approximate a straight line. Consider transforming the continuous variable if this is not the case. Another method to check this property of continuous covariates uses fractional polynomials.

Another method is to use two residual-based plots: (1) a plot of the covariate values versus the Martingale residuals (and their smooth) from a model that excludes the covariate of interest, and (2) a plot of the covariate values versus the log of the ratio of smoothed censor to smoothed cumulative hazard. To construct the second plot: (1) fit the preliminary main effects model, including the covariate of interest (e.g. 'age'), (2) save the Martingale residuals (M_i) from this model, (3) calculate $H_i = c_i - M_i$, where c_i is the censoring variable, (4) plot the values of c_i versus the covariate of interest and calculate a lowess smooth (called c_{LSM}), (5) plot the values of H_i versus the covariate of interest and calculate a lowess smooth (called H_{LSM}), (6) the smoothed values from these plots are used to calculate:

$$y_i = ln\left(\frac{c_{LSM}}{H_{LSM}}\right) + \beta_{age} \times age_i \tag{8}$$

and the pairs (y_i, age_i) are plotted and connected by straight lines. There should be a linear relationship between the covariate values and each of the described parameters.

Interactions

The final step is to determine whether interaction terms are required. An interaction term is a new variable that is the product of two other variables in the model. Note that there can be subject matter considerations that dictate that a particular interaction term (or terms) should be included in a given model, regardless of their statistical significance. In most settings there is no biological or clinical theory to justify automatic inclusion of interactions.

The effect of adding an interaction term should be assessed using the partial likelihood ratio test. All significant interactions should be included in the main-effects model. Wald statistic P-values can be used as a guide to selecting interactions that may be eliminated from the model, with significance checked by the partial likelihood ratio test.

At this point we have a 'preliminary model' and the next step is to assess its fit and adherence to key assumptions.

5.2   Testing the proportional hazards assumption

Once a suitable set of covariates has been identified, it is wise to check each covariate to ensure that the proportional hazards assumption is valid. To assess the proportional hazards assumption we examine the

extent to which the estimated hazard curves for each level of strata of a covariate are equidistant over time.

A plot of the scaled Schoenfeld residuals (and a loess smoother) as a function of time and the partial likelihood ratio test may be used to test proportionality of hazards. In a 'well-behaved' model the Schoenfeld residuals are scattered around 0 and a regression line fitted to the residuals has a slope of approximately 0. The fitted curves can be interpreted as estimates of $\beta(t)$ if the coefficients are allowed to vary over time.

For categorical covariates the proportional hazards assumption can be visually tested by plotting -log[-log $S(t)$] vs time for strata of each covariate. If the proportionality assumption holds the two (or more) curves should be approximately parallel and should not cross. Alternatively, run a model with each covariate (individually). Introduce a time-dependent interaction term for that covariate. If the proportional hazards assumption is valid for the covariate, the introduction of the time-dependent interaction term won't be significant. This approach is regarded as the most sensitive (and objective) method for testing the proportional hazards assumption.

What do you do if a covariate violates the proportional hazards assumption? The first option is to stratify the model by the offending covariate. This means that a separate baseline hazard function is produced for each level of the covariate. Note you can't obtain a hazard ratio for the covariate you've stratified on because its influence on survival is 'absorbed' into the (two or more) baseline hazard functions in the stratified model. If you are interested in quantifying the effect of the covariate on survival then you should introduce a time-dependent interaction term for the covariate, as described above.

5.3   Residuals

Residuals analysis provide information for evaluating a fitted proportional hazards model. They identify leverage and influence measures and can be used to assess the proportional hazards assumption. By definition, residuals for censored observations are negative and residual plots are useful to get a feeling for the amount of censoring in the data set — large amounts of censoring will result in 'banding' of the residual points. There are three types of residuals:

1. Martingale residuals. Martingale residuals are the difference between the observed number of events for an individual and the conditionally expected number given the fitted model, follow up time, and the observed course of any time-varying covariates. Martingale residuals may be plotted against covariates to detect non-linearity (that is, an incorrectly specified functional form in the parametric part of the model). Martingale residuals are sometimes referred to as Cox-Snell or modified Cox-Snell residuals.

2. Score residuals. Score residuals should be thought of as a three-way array with dimensions of subject, covariate and time. Score residuals are useful for assessing individual influence and for robust variance estimation.

3. Schoenfeld residuals. Schoenfeld residuals are useful for assessing proportional hazards. Schoenfeld residuals provide greater diagnostic power than unscaled residuals. Sometimes referred to as score residuals.

5.4   Overall goodness-of-fit

To assess the overall goodness-of-fit of a Cox proportional hazards regression model Arjas (1988) suggests plotting the cumulative observed versus the cumulative expected number of events for subjects with observed (not censored) survival times. If the model fit is adequate, then the points should follow a 45°

line beginning at the origin. The methodology is as follows: (1) create groups based on covariate values (e.g. treated yes, treated no) and sort on survival time within each group, (2) compute the cumulative sum of the zero-one censoring variable and the cumulative sum of the the cumulative hazard function within each group, (3) plot the pairs of cumulative sums within each group only for subjects with an observed survival time.

As in all regression analyses some sort of measure analogous to R^2 may be of interest. Schemper and Stare (1996) show that there is not a single simple, easy to calculate, easy-to-interpret measure to assess the goodness-of-fit of a proportional hazards regression model. Often, a perfectly adequate model may have what, at face value, seems like a very low R^2 due to a large amount of censoring. Hosmer and Lemeshow recommend the following as a summary statistic for goodness of fit:

$$R_M^2 = 1 - exp\left[\frac{2}{n}(L_0 - L_M)\right] \tag{9}$$

Where:

L_0: the log partial likelihood for the intercept-only model,
L_M: the log partial likelihood for the fitted model,
n: the number of cases included.

5.5 Worked examples

Selection of covariates

```
Load the survival library. Read the addict data file into R:

library(survival)
setwd("D:\\TEMP")
dat <- read.table("addict.csv", header = TRUE, sep = ",")

Set contrasts for clinic and prison. Set the reference category for clinic, making Clinic 1 (base = 1) the reference category:

dat$clinic <- factor(dat$clinic, levels = c(1, 2), labels = c("1", "2"))
contrasts(dat$clinic) <- contr.treatment(2, base = 1, contrasts = TRUE)
levels(dat$clinic)

Same for prison, making absence of a prison record the reference category:

dat$prison <- factor(dat$prison, levels = c(0, 1), labels = c("0", "1"))
contrasts(dat$prison) <- contr.treatment(2, base = 1, contrasts = TRUE)
levels(dat$prison)

Assess the influence of clinic, prison and dose on days to relapse. First of all categorise dose into four classes based on quartiles:

quantile(dat$dose, probs = c(0.25, 0.50, 0.75))
hist(dat$dose)

Quartiles for dose are 50, 60 and 70. Create a categorical variable based on dose:

dose.cat <- rep(0, length(dat[,1]))
dose.cat[dat$dose < 50] <- 1
dose.cat[dat$dose >= 50 & dat$dose < 60] <- 2
dose.cat[dat$dose >= 60 & dat$dose < 70] <- 3
dose.cat[dat$dose >= 70] <- 4
dat <- cbind(dat, dose.cat)
```

Assess the effect of `clinic`, `prison` and `dose.cat` on days to relapse:

```
addict.km01 <- survfit(Surv(stop, status) ~ clinic, type = "kaplan-meier",
data = dat)
addict.km02 <- survfit(Surv(stop, status) ~ prison, type = "kaplan-meier",
data = dat)
addict.km03 <- survfit(Surv(stop, status) ~ dose.cat, type = "kaplan-meier",
data = dat)
```

Plot all Kaplan-Meier curves on one page. The `mark.time = FALSE` argument disables the censor marks:

```
par(pty = "s", mfrow = c(2,2))
plot(addict.km01, xlab = "Days to relapse", ylab = "Cumulative proportion to
experience event", main = "Clinic", lty = c(1,2), mark.time = FALSE,
legend.text = c("Clinic 1","Clinic 2"), legend.pos = 0, legend.bty = "n", cex
= 0.80)
plot(addict.km02, xlab = "Days to relapse", ylab = "Cumulative proportion to
experience event", main = "Prison", lty = c(1,2), mark.time = FALSE,
legend.text = c("Prison absent","Prison present"), legend.pos = 0, legend.bty
= "n", cex = 0.80)
plot(addict.km03, xlab = "Days to relapse", ylab = "Cumulative proportion to
experience event", main = "Dose categories", lty = c(1,2,3,4), mark.time =
FALSE, legend.text = c("Dose 1","Dose 2", "Dose 3", "Dose 4"), legend.pos = 0,
legend.bty = "n", cex = 0.80)
```

Log-rank tests:

```
survdiff(Surv(stop, status) ~ clinic, data = dat, na.action = na.omit, rho =
0)
survdiff(Surv(stop, status) ~ prison, data = dat, na.action = na.omit, rho =
0)
survdiff(Surv(stop, status) ~ dose.cat, data = dat, na.action = na.omit, rho =
0)
```

The variables `clinic` and dose (as a categorical variable) influence days to relapse. The variable `prison` is not significant when tested with a log-rank test ($P = 0.28$), but since it is considered to be biologically important it is retained in our model.

Fit multivariable model

Days to relapse depends on `clinic`, `prison` and dose:

```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose, method = "breslow", data = dat)
summary(addict.cph01)
```

Variables `clinic` and dose significantly influence time to relapse ($P = 2.6E-06$ and $3.1E-08$, respectively). Variable `prison` approaching significance ($P = 0.06$). Drop variable `prison` (using the update function):

```
addict.cph02 <- update(addict.cph01, ~. - prison)
summary(addict.cph02)
```

Does `addict.cph02` provide a better fit to the data than `addict.cph01`?

```
x2 <- 2 * (addict.cph02$loglik[2] - addict.cph01$loglik[2])
1 - pchisq(x2,1)
```

Removing prison has no effect on model fit ($P = 1.0$). Since it is felt that the presence of a prison record is important in this context it is retained in the model:

```
summary(addict.cph01)
```

Check scale of continuous covariates (method 1)

Replace the continuous covariate dose with design (dummy) variables. Plot the estimated coefficients versus the midpoint of each group:

```
dat$clinic <- factor(dat$clinic, levels = c(1, 2), labels = c("1", "2"))
contrasts(dat$clinic) <- contr.treatment(2, base = 1, contrasts = TRUE)
dat$prison <- factor(dat$prison, levels = c(0, 1), labels = c("0", "1"))
contrasts(dat$prison) <- contr.treatment(2, base = 1, contrasts = TRUE)

dose.cat <- rep(0, length(dat[,1]))
dose.cat[dat$dose < 50] <- 1
dose.cat[dat$dose >= 50 & dat$dose < 60] <- 2
dose.cat[dat$dose >= 60 & dat$dose < 70] <- 3
dose.cat[dat$dose >= 70] <- 4
dat <- cbind(dat, dose.cat)

dat$dose.cat <- factor(dat$dose.cat, labels=c("1", "2", "3", "4"))
contrasts(dat$dose.cat) <- contr.treatment(4, base = 1, contrasts = TRUE)

addict.cph03 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose.cat, method = "breslow", data = dat)
summary(addict.cph03)

addict.cph03$coefficients
addict.cph03$coefficients[3:5]

x <- c(((50 + min(dat$dose))/2), 55, 65,((max(dat$dose) + 70)/2))
y <- c(0, addict.cph03$coefficients[3:5])
plot(x, y, xlim = c(0, 100), type = "l", xlab = "Dose", ylab = "Regression
coefficient")
```

Scale of continuous covariates linear — no transformations required for variable dose.

Check scale of continuous covariates (method 2)

```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose, method = "breslow", data = dat)
addict.mi <- residuals(addict.cph01, type = "martingale")
addict.hi <- dat$status - addict.mi
addict.clsm <- lowess(dat$dose, dat$status)
addict.hlsm <- lowess(dat$dose, addict.hi)
addict.yi <- log(addict.clsm$y / addict.hlsm$y) +
(addict.cph01$coefficients[3] * dat$dose)
```

Plot covariate values versus Martingale residuals and `addict.yi` versus covariate values:

```
par(pty = "s", mfrow = c(1,2))
plot(dat$dose, addict.mg)
lines(lowess(dat$dose, addict.mg))
plot(addict.yi, dat$dose)
```

A linear relationship is evident between the covariate valuesand each of the calculated parameters, indicating that described the continuous variable dose is linear in its log hazard.

Interactions

Check for significance of the interaction between the categorical variables `clinic` and `prison`:

```
addict.cph04 <- coxph(Surv(stop, status, type="right") ~ clinic + prison +
dose + (clinic * prison), method = "breslow", data = dat)
summary(addict.cph04)
```

The P value of the Wald test for the interaction term `clinic * prison` is not significant (P = 0.11).

Testing the proportional hazards assumption

Plot scaled Schoenfeld residual plots:

```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose, method = "breslow", data = dat)
addict.zph <- cox.zph(addict.cph01)

par(pty = "s", mfrow = c(2,2))
plot(addict.zph[1], main = "Clinic"); abline(h = 0, lty = 2)
plot(addict.zph[2], main = "Prison"); abline(h = 0, lty = 2)
plot(addict.zph[3], main = "Dose"); abline(h = 0, lty = 2)
```

The variability band for `clinic` displays a negative slope over time, suggesting non-proportionality of hazards. Formally test the proportional hazards assumption for all variables in `addict.cph01`:

```
cox.zph(addict.cph01, global = TRUE)
```

Using the `cox.zph` function, `rho` is the Pearson product-moment correlation between the scaled Schoenfeld residuals and time. The hypothesis of no correlation is tested using test statistic `chisq`. In the above example, the significant `cox.zph` test for `clinic` (P < 0.01) implies that the proportional hazards assumption as been violated for the `clinic` variable. This notion is supported by the Schoenfeld residual plots. An alternative (and less sensitive) means of testing the proportional hazards assumption is to plot log[-log $S(t)$] vs time:

```
clinic.km <- survfit(Surv(stop, status) ~ clinic, type = "kaplan-meier", data
= dat)
clinic.km$strata[1]
clinic <- c(rep(1, times = clinic.km$strata[1]), rep(2, times =
clinic.km$strata[2]))
clinic.haz <- as.data.frame(cbind(clinic, time = clinic.km$time, surv =
clinic.km$surv))
clinic1 <- log(-log(clinic.haz$surv[clinic.haz$clinic == 1]))
clinic2 <- log(-log(clinic.haz$surv[clinic.haz$clinic == 2]))

plot(c(clinic.haz$time[clinic.haz$clinic == 1],
clinic.haz$time[clinic.haz$clinic == 2]), c(clinic1, clinic2), type = "n",
ylim = c(-5, 2), xlab = "Days to relapse", ylab = "Log cumulative hazard",
main = "Clinic")
lines(clinic.haz$time[clinic.haz$clinic == 1], clinic1, type = "s", lty = 1)
lines(clinic.haz$time[clinic.haz$clinic == 2], clinic2, type = "s", lty = 2)
legend(x = "topleft", legend = c("Clinic 1", "Clinic 2"), lty = c(1, 2), bty =
"n")
```

We could be talked into concluding that the -log[-log $S(t)$] vs time plots for `clinic` are parallel — conflicting with the findings of the `cox.zph` test and the Schoenfeld residual plots.

Residuals

Deviance residuals:

```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose, method = "breslow", data = dat)
addict.res <- residuals(addict.cph01, type = "deviance")

par(pty = "s", mfrow = c(2, 2))
boxplot(addict.res ~ dat$clinic, main = "Clinic"); abline(h = 0, lty = 2)
boxplot(addict.res ~ dat$prison, main = "Prison"); abline(h = 0, lty = 2)
plot(dat$dose, addict.res, xlab = "Dose", ylab = "Deviance residual", main =
"Dose"); abline(h = 0, lty = 2)
```

The following plots show the change in each regression coefficient when each observation is removed from the data (influence statistics). The changes plotted are scaled in units of standard errors and changes of less than 0.1 are of little concern. Plot influence statistics (using a common scale for the vertical axis: -0.1 to +0.1):

```
addict.res <- resid(addict.cph01, type = "dfbeta")
par(mfrow = c(2, 2))
main <- c("Clinic", "Prison", "Dose")
for (i in 1:3){
plot(1:238, addict.res[,i], type = "h", ylim = c(-0.1,0.1), xlab =
"Observation", ylab = "Change in coefficient")
    title(main[i])
  }
```

The above plots give an idea of the influence individual observations have on the estimated regression coefficients for each covariate. Data sets where the influence plot is tightly clustered around zero indicate an absence of influential observations. Now plot the Martingale residuals:

```
res <- residuals(addict.cph01, type = "martingale")
X <- as.matrix(dat[,c("clinic", "prison", "dose")])

par(mfrow = c(2,2))
for(j in 1:3){
plot(X[,j], res, xlab = c("Clinic", "Prison", "Dose")[j], ylab = "Martingale
residuals")
  abline(h = 0, lty = 2)
  lines(lowess(X[,j], res))
  }

par(mfrow = c(2,2))
b <- coef(addict.cph01[1:3])
for(j in 1:3){
  plot(X[,j], b[j] * X[,j] + res, xlab = c("Clinic", "Prison", "Dose")[j],
    ylab = "Component + residual")
  abline(lm(b[j] * X[,j] + res ~ X[,j]), lty = 2)
  lines(lowess(X[,j], b[j] * X[,j] + res, iter = 0))
  }
```

Overall goodness of fit

```
Cox model:

addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose, method = "breslow", data = dat)
summary(addict.cph01)

Log partial likelihood for the [intercept-only] model and for the fitted model:

addict.cph01$loglik[1]; addict.cph01$loglik[2]

Compute Schemper and Stare (1996) R² manually:

r.square <- 1 - exp( (2/length(dat[,1])) * (addict.cph01$loglik[1] -
addict.cph01$loglik[2]))
r.square
```

Dealing with violation of the proportional hazards assumption

From the analyses conducted so far, we conclude that the proportional hazards assumption has been violated for the variable `clinic`. One method of dealing with this is to stratify the model by `clinic`. This means that we produce a separate baseline hazard function for each level of `clinic`. Note that by stratifying we cannot obtain a hazard ratio for `clinic` since the 'clinic effect' is absorbed into the baseline hazard.

```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose, method = "breslow", data = dat)
addict.cph04 <- coxph(Surv(stop, status, type = "right") ~ strata(clinic) +
prison + dose, method = "breslow", data = dat)
summary(addict.cph04)
```

Compare the `clinic + dose + prison` model with the stratified model:

```
x2 <- 2 * (addict.cph04$loglik[2] - addict.cph01$loglik[2])
1 - pchisq(x2, 1)
```

The stratified model provides a significantly better fit. Parameterising `clinic` as a time dependent covariate would be one option for dealing with non-proportionality of hazards and retaining the ability to quantify the effect of clinic. Plot Kaplan-Meier survival curves for each clinic, adjusting for the effect of `prison` and methadone dose:

```
plot(survfit(addict.cph04), lty = c(1,3), xlab = "Days to relapse", ylab =
"Cumulative proportion to experience event", legend.text = c("Clinic
1","Clinic 2"), legend.pos = 0, legend.bty = "n")
```

5.6 Poisson regression

A Cox model can be estimated using standard Poisson regression techniques by splitting the data finely and specifying the model as having one rate parameter per time interval. The relevant time-varying covariates should be computed for each interval and fixed covariates should be carried over to all intervals for a given individual. The split data makes a clear destinction between risk time which is the length of each time interval and time scale which is the value of the time scale at (the beginning of) each interval. Using a Poisson approach the log-risk time is used as offset and a smoothed estimate of time is used as a covariate.

Since everything possible using a Cox approach can be done using Poisson modelling of split data, there is no loss of capability by switching to Poisson modelling. When stratification or time-dependent variables are required, the facilities using a standard Cox approach limits the ways in which the desired interactions can be modelled, and can distract the analyst from realising that other interactions between covariates may be of interest.

6   Parametric models

As discussed, semi-parametric models make no assumption about the distribution of failure times, but do make assumptions about how covariates change survival experience. Parametric models, on the other hand, make assumptions about the distribution of failure times and the relationship between covariates and survival experience. Parametric models fully specify the distribution of the baseline hazard/survival function according to some (defined) probability distribution. Parametric models are useful when we want to predict survival rather than identify factors that influence survival. Parametric models can be expressed in: (1) proportional hazard form, where a one unit change in an explanatory variable causes proportional changes in hazard; and (2) accelerated failure time (AFT) form, where a one unit change in an explanatory variable causes a proportional change in survival time. The advantage of the accelerated failure time approach is that the effect of covariates on survival can be described in absolute terms (e.g. numbers of years) rather than relative terms (a hazard ratio).

6.1   Exponential model

The exponential model is the simplest type of parametric model in that it assumes that the baseline hazard is constant over time:

$$h(t) = h_0 \exp^{\beta X} \text{ where } h_0 = \lambda \tag{10}$$

The assumption that the baseline hazard is constant over time can be evaluated in several ways. The first method is to generate an estimate of the baseline hazard from a Cox proportional hazards model and plot it to check if it follows a straight, horizontal line. A second approach is to fit a model with a piecewise-constant baseline hazard. Here, the baseline hazard is allowed to vary across time intervals (by including indicator variables for each of the time intervals). The baseline hazard is assumed to be constant within each time period, but can vary between time periods.

6.2   Weibull model

In a Weibull model it is assumed that the baseline hazard has a shape which gives rise to a Weibull distribution of survival times:

$$h(t) = h_0 \exp^{\beta X} \text{ where } h_0 = \lambda p t^{p-1} \tag{11}$$

Where βX includes an intercept term β_0. The suitability of the assumption that survival times follow a Weibull distribution can be assessed by generating a log-cumulative hazard plot. If the distribution is Weibull, this function will follow a straight line. The estimated shape parameter from the Weibull model gives an indication of whether hazard is falling ($p < 1$), constant ($p = 1$), or increasing ($p > 1$) over time.

6.3   Accelerated failure time models

The general form of an accelerated failure time model is:

$$log(t) = \beta X + log(\tau) \text{ or } t = \exp^{\beta X} \tau \tag{12}$$

where $log(t)$ is the natural log of the time to failure event, βX is a linear combination of explanatory variables and $\log{(\tau)}$ is an error term. Using this approach τ is the distribution of survival times when

Table 5: Accelerated failure time model showing the effect of clinic, methadone dose and prison status on expected retention time on the program (adapted from Caplehorn and Bell 1991). Note that the term 'hazard' in the last column of the table is replaced with 'survival.'

Variable	Subjects	Failed	Coefficient (SE)	P	Survival (95% CI)
Intercept	238	250	4.7915 (0.2782)	< 0.01	
Clinic				< 0.01	
Clinic 1	163	122			1.0
Clinic 2	75	28	0.7198 (0.1595)		2.05 (1.50 − 2.81) [b]
Prison record				0.07	
Absent	127	81			1.0
Present	111	69	-0.2232 (0.1224)		0.80 (0.63 − 1.02)
Dose	238	150	0.0247 (0.0046)	< 0.01	1.02 (1.01 − 1.03)

[a] The significance of inclusion of the two clinic variables in the model.
[b] Intepretation: after adjusting for the effect of methadone dose and prison status retention time for patients from Clinic 2 was twice that of patients from Clinic 1 (95% CI 1.50 − 2.81).
SE: standard error.
CI: confidence interval.

$\beta X = 0$. If we assume that τ follows a log-normal distribution, then the log of survival times will have a normal distribution, which is equivalent to fitting a linear model to the natural log of survival time (assuming that you can ignore the problem of dealing with censored observations). Equation 12 can be re-expressed as follows:

$$\tau = \exp^{-\beta X} t \text{ or } ln(\tau) = -\beta X + log(t) \tag{13}$$

The linear combination of predictors in the model (βX) can act additively or multiplicatively on the log of time: they speed up or slow down time to event by a multiplicative factor. In this case $\exp^{-\beta X}$ is called the acceleration parameter such that if $\exp^{-\beta X} > 1$ time passes more quickly, if $\exp^{-\beta X} = 1$ time passes at a normal rate, and if $\exp^{-\beta X} < 1$ time passes more slowly.

Exponential and Weibull models can be parameterised as either proportional hazards models or as accelerated failure time models. Other parametric models (e.g. the log-normal, the log-logistic, and gamma) can only be expressed as accelerated failure time models (the predictors in these models do not necessarily multiply the baseline hazard by a constant amount).

Accelerated failure time coefficients represent the expected change in $ln(t)$ for a one unit change in the predictor. Consider an accelerated failiure time model fitted to the `addict` data, as shown in Table 5. What was the effect of being treated at Clinic 2 in terms of additional retention time?

$log(t) = 4.7915 + (0.7198 \times 1)$
$log(t) = 5.5113$
$t = \exp(5.5113)$
$t = 247$ days

Being treated at Clinic 2 extended retention time by 247 days.

6.4 Worked examples

Exponential and Weibull models

```
library(survival)
setwd("D:\\TEMP")
dat <- read.table("addict.csv", header = TRUE, sep = ",")
```

Set contrasts for clinic and prison. Set the reference category for `clinic`, making Clinic 1 (base = 1) the reference category:

```
dat$clinic <- factor(dat$clinic, levels = c(1, 2), labels = c("1", "2"))
contrasts(dat$clinic) <- contr.treatment(2, base = 1, contrasts = TRUE)
levels(dat$clinic)
```

Same for prison, making absence of a prison record the reference category:

```
dat$prison <- factor(dat$prison, levels = c(0, 1), labels = c("0", "1"))
contrasts(dat$prison) <- contr.treatment(2, base = 1, contrasts = TRUE)
levels(dat$prison)
```

Cox proportional hazards model (for comparison):

```
addict.cph01 <- coxph(Surv(stop, status, type = "right") ~ clinic + prison +
dose, method = "breslow", data = dat)
summary(addict.cph01)
```

Exponential model:

```
addict.exp01 <- survreg(Surv(stop, status, type = "right") ~ clinic + prison +
dose, dist = "exp", data = dat)
summary(addict.exp01)

shape.exp = 1/ addict.exp01$scale
shape.exp
```

Weibull model:

```
addict.wei01 = survreg(Surv(stop, status, type = "right") ~ clinic + prison +
dose, dist = "weib", data = dat)
summary(addict.wei01)

shape.wei = 1 / addict.wei01$scale
shape.wei
```

Compare the three models using AIC:

```
extractAIC(addict.cph01)
extractAIC(addict.exp01)
extractAIC(addict.wei01)
```

The AIC for the Cox model is the smallest, indicating that this model provides the best fit with the data (this is consistent with the diagnostics we ran earlier to assess how consistent the data was with the exponential and Weibull distributions). Additional plotting options are available using the `Design` package. Re-run Weibull model using the psm function:

```
library(Design)
addict.wei <- psm(Surv(stop, status) ~ clinic + prison + dose, dist = "weib",
data = dat)
```

Plot survivorship for each clinic for a patient without a prison record and receiving a maximum dose of 50 mg methadone per day:

```
survplot(addict.wei, what = "survival", clinic = c("1","2"), prison = "0",
dose = 50)
```

Cumulative hazard:

```
survplot(addict.wei, what = "hazard", clinic = c("1","2"), prison = "0", dose
= 50)
```

Accelerated failure time models

Here we use the psm function in the `Design` library to develop an AFT model. The psm function is a modification of `survreg` and is used for fitting the accelerated failure time family of parametric survival models.

```
addict.aft01 <- psm(Surv(stop, status) ~ clinic + prison + dose, dist =
"weib", data = dat)
addict.aft01
```

What is the effect of Clinic 2 on retention time (after adjusting for the effect of presence of a prison record and maximum daily methadone dose)?

```
exp(addict.aft01$coefficients[2])
```

Treatment at Clinic 2 doubles patient retention time. What does this mean in terms of calendar time?

```
log.t <- as.numeric(addict.aft01$coefficients[1] +
(addict.aft01$coefficients[2]* 1))
exp(log.t)
```

Treatment at Clinic 2 results in patients remained on the program for an additional 247 days (compared with those treated at Clinic 1).

7 Time dependent covariates

As discussed, stratification as a method for dealing with a covariate that violates the proportional hazards assumption is not an option when you want to include it in a model, in order to describe and/or test its effect on survival. In this case it can be useful to look at a plot of the hazard function versus time for different strata of the variable. This may indicate the type of deviation from proportional hazards that is occurring. Two common types of departure from proportional hazards in clinical situations are: (1) the time to peak hazard varies between prognostic groups, i.e. strata, or (2) the influence of a covariate diminishes with time.

If the covariate is fixed (i.e. it does not vary with time, but its effect appears to vary with time) we can explore this time-dependent effect by dividing the time period into distinct intervals. We then fit proportional hazards models to the survival in each interval and compare the coefficients for each covariate across the different time intervals. If the coefficient changes with time, we have evidence of non-proportional hazards. This approach has been called the step function proportional hazards or piecewise Cox model. In each interval, patients who die or who are censored before the interval are [treated as usual - i.e. coded as censored or died] and survival times for patients who live through the interval to the next one are censored at the end of the interval. The number and length of the intervals is arbitrary, but each interval should contain enough deaths to enable regression coefficients to be estimated reliably.

Covariates themselves may change with time. In this case the survival period for each individual is divided up into a sequence of shorter 'survival spells', each characterised by an entry and an exit time, and within which covariate values remain fixed. Thus the data for each individual are represented by a number of shorted censored intervals and possibly one interval ending with the event of interest (death, for example).

It may be thought that the observations, when organised in this way, are 'correlated' and so not suitable for Cox regression. Fortunately, this is not an issue, since the partial likelihood on which estimation is based has a term for each unique death or event time, and involves sums over those observations that are available or at risk at the actual event date. Since the intervals for a particular individual do not overlap, the likelihood will involve at most only one of the observations for the individual, and so will be based on independent observations. The values of the covariates between event times do not enter the partial likelihood.

References

Black, D., & French, N. (2004). Effects of three types of trace element supplementation on the fertility of three commercial dairy herds. *Veterinary Record*, *154*, 652 - 658.

Caplehorn, J., & Bell, J. (1991). Methadone dosage and retention of patients in maintenance treatment. *Medical Journal of Australia*, *154*(3), 195 - 199.

Collett, D. (1994). *Modelling Survival Data in Medical Research*. London: Chapman and Hall.

Dohoo, I., Martin, S., & Stryhn, H. (2003). *Veterinary Epidemiologic Research*. Charlottetown, Prince Edward Island, Canada: AVC Inc.

Fisher, L., & Lin, D. (1999). Time-dependent covariates in the Cox proportional hazards regression model. *Annual Reviews in Public Health*, *20*, 145 - 157.

Haerting, J., Mansmann, U., & Duchateau, L. (2007). *Frailty Models in Survival Analysis*. Unpublished doctoral dissertation, Martin-Luther-Universität Halle-Wittenberg.

Kleinbaum, D. (1996). *Survival Analysis: A Self-Learning Text*. New York: Springer-Verlag.

Lee, E. (1992). *Statistical Methods for Survival Analysis*. London: Jon Wiley and Sons Inc.

Lee, E., & Go, O. (1997). Survival analysis in public health research. *Annual Reviews in Public Health*, *18*, 105 - 134.

Leung, K., Elashoff, R., & Afifi, A. (1997). Censoring issues in survival analysis. *Annual Reviews in Public Health*, *18*, 83 - 104.

More, S. (1996). The performance of farmed ostrich eggs in eastern Australia. *Preventive Veterinary Medicine*, *29*, 121 - 134.

Proudman, C., Dugdale, A., Senior, J., Edwards, G., Smith, J., Leuwer, M., et al. (2006). Pre-operative and anaesthesia-related risk factors for mortality in equine colic cases. *The Veterinary Journal*, *171*(1), 89 - 97.

Proudman, C., Pinchbeck, G., Clegg, P., & French, N. (2004). Risk of horses falling in the Grand National. *Nature*, *428*, 385 - 386.

Stevenson, M., Wilesmith, J., Ryan, J., Morris, R., Lockhart, J., Lin, D., et al. (2000). Temporal aspects of the bovine spongiform encephalopathy epidemic in Great Britain: Individual animal-associated risk factors for disease. *Veterinary Record*, *147*(13), 349 - 354.

Tableman, M., & Kim, J. (2004). *Survival Analysis Using S*. New York: Chapman Hall/CRC.

The Diabetes Control and Complications Trial Research Group. (1996). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *New England Journal Of Medicine*, *329*(14), 977 - 986.

Therneau, T., & Grambsch, P. (2001). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.

Venables, W., & Ripley, B. (2002). *Modern Applied Statistics with S*. New York: Springer-Verlag.

Wilesmith, J., Ryan, J., Stevenson, M., Morris, R., Pfeiffer, D., Lin, D., et al. (2000). Temporal aspects of the bovine spongiform encephalopathy epidemic in Great Britain: Holding-associated risk factors for disease. *Veterinary Record*, *147*(12), 319 - 325.

Wilesmith, J., Stevenson, M., King, C., & Morris, R. (2003). Spatio-temporal epidemiology of foot-and-mouth disease in two counties of Great Britain in 2001. *Preventive Veterinary Medicine*, *61*(3), 157 - 170.