AUSTRALIAN NATIONAL UNIVERSITY

SCHOOL OF FINANCE AND APPLIED STATISTICS

SURVIVAL MODELS 1 (STAT3032/STAT8042)
BIOSTATISTICS (STAT8003)

Part 1

Sections 1 to 4

1. Estimation of the survival distribution

1.1 Introduction

There are many different ways of quantifying the random behaviour of the lifetime T of an individual where T is assumed to have a continuous distribution. Recall that some of the possibilities are:

- the survival distribution (sdf), $S_T(t) = \Pr(T > t)$, which is the probability that the individual lives past t,
- the cumulative distribution function (cdf) $F_T(t) = \Pr(T \leq t)$, which is the probability that the individual dies by t,
- the probability density function (pdf), $f_T(t) = -S'_T(t) = F'_T(t)$, where $f_T(t)d$ is approximately the probability of dying in the interval $(t, t+d)$,
- the hazard function, $m_t = m_T(t) = -S'_T(t)/S_T(t) = F'_T(t)/S_T(t) = f_T(t)/S_T(t)$, where $m_T(t)d$ is approximately the probability of dying in the interval $(t, t+d)$ given survival until t. It is also common to use the notation $l_T(t)$ for the hazard function.
- the integrated hazard function, $\Lambda_T(t) = \int_0^t l_T(y)dy$. Heuristically this is the amount of hazard that an individual has accumulated by time t and the sdf is $S_T(t) = \exp(-\Lambda_T(t))$.

It is largely a matter of convenience which function is used and any of the functions can be used to derive the others. For example, the fundamental connection between the hazard function and the survival distribution function is

$$S_T(t) = \exp\left(-\int_0^t m_T(y)dy \right). \qquad (1.1)$$

Example 1.1: Exponential

The exponential distribution has constant hazard function $m_T(t) = l$. Using the fundamental relationship (1.1), it follows that the associated sdf is $S_T(t) = \exp(-lt)$, the cdf is $F_T(t) = 1 - \exp(-lt)$ and the pdf is $f_T(t) = l\exp(-lt)$. Note that the assumption of constant hazard at all ages is very unlikely to hold in practice across the whole lifetime of the individual. We do not expect that old people will have the same probability of dying in the next year as young people. At best, the exponential assumption will hold for small portions of the lifetime of an individual.

Two approaches to estimating the sdf are possible. One assumes a specific family for the sdf, for example exponential, and then estimates the parameters of that family. A second approach, to be considered in this section, does not assume a specific model. Instead a non-parametric approach is used. An estimate is constructed which does not depend on specific assumptions concerning the underlying sdf and which has desirable properties in general.

1.2 Empirical Distribution Function

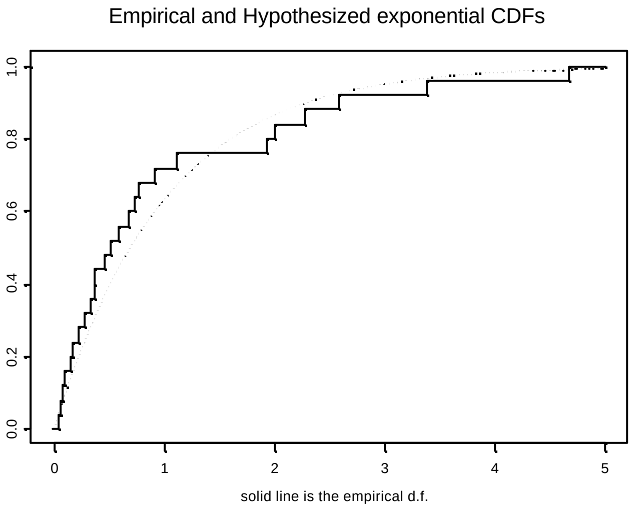
Suppose it is possible to observe the lifetimes of a set of N individuals. Then the obvious estimate of the probability of dying by age t, $F_T(t) = \Pr(T \le t)$, is $\hat{F}_T(t) = d(t)/N$ where $d(t)$ is the numbers of individuals who have died by time t. The function $\hat{F}_T(t)$ for $0 \le t < \infty$ is called the empirical distribution function. Since $d(t)$ is a Binomial random variable with parameters N and $F_T(t)$, it follows that $\hat{F}_T(t)$ is unbiased for $F_T(t)$ and has variance $F_T(t)(1-F_T(t))/N$. Since $\hat{F}_T(t)$ is a function, it also makes sense to consider the covariance and correlation of the estimate at two different time points. We will see in the first tutorial that for $s \le t$, we can show that

$$Cov\left(\hat{F}_T(s), \hat{F}_T(t)\right) = F_T(s)(1-F_T(t))/N .$$

Example 1.2

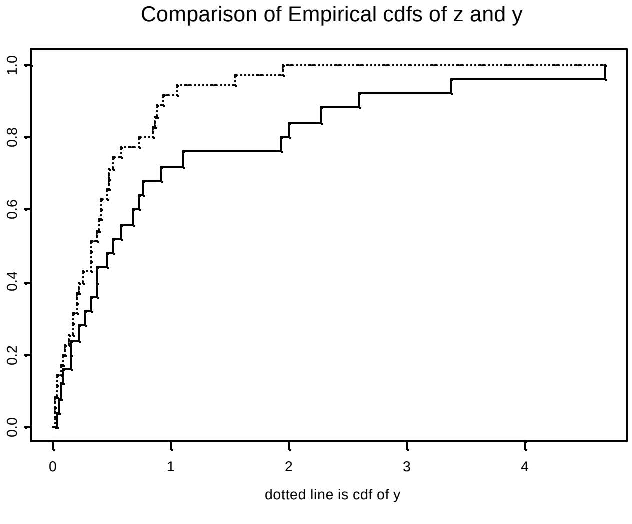
We will use Splus to see what an empirical cdf looks like. A convenient function is cdf.compare( ) which will calculate one or two empirical cdf's. Suppose we generate 25 observations from a standard exponential and compare the empirical cdf to the cdf of the standard exponential. The results are:

```
> z_rexp(25)
> z
 [1] 0.68413871 0.28335138 0.73550926 0.09825028
 [5] 0.15276740 2.01282458 0.45908951 0.32519581
 [9] 4.67691654 0.06306868 1.93365203 0.37809371
[13] 0.22391922 0.07703267 0.50979374 0.57870186
[17] 0.16827660 2.59678869 2.28199860 3.38493919
[21] 1.10714227 0.03517625 0.92107975 0.37357329
[25] 0.76561069
> cdf.compare(z,dist="exp")
```



We can also visually compare two samples:

```
> y_rexp(35,rate=2)
> y
 [1] 0.17892730 0.20224836 0.26092400 0.48429994
 [5] 0.18248288 0.86952302 0.32420025 0.58079872
 [9] 0.88516342 1.55566225 0.03226694 0.85957289
[13] 0.74186042 0.47031661 0.39127159 0.07984834
[17] 0.42095765 0.42159099 0.02004534 0.08605680
[21] 0.37738644 0.13884174 0.21381211 1.05458276
[25] 0.52259665 0.32376906 0.94208211 0.47494508
[29] 0.03646690 1.95693579 0.10852609 0.04745564
[33] 0.22636855 0.33272190 0.01639930
> cdf.compare(z,y)
```



In many practical situations, it is unrealistic to expect that the actual lifetimes of individuals can be observed. For example if we wished to look at the effect of taking aspirin on the survival times of heart attack patients, it may not be possible (or desirable) to wait until all patients have died since that may take many years. It is often very important to get answers to medical questions in a shorter time frame. Information is often collected and analysed before all individuals have died. If individuals have not died, then they are called censored data. One approach to handling censored data is to simply discard the data points. This can lead to bias since individuals with long lifetimes are discarded. Also, even though the exact time of death is not known, we have the information that the time of death is later than the time of censoring which should be used if possible in constructing the estimate of the sdf. There are several different forms of censoring which are common in practice which will be considered in the next section.

1.3 Censoring Mechanisms

Data are said to be censored if we know that the data point belongs to a certain interval, but we do not know the exact value. For example if when we terminated a study, a patient of age 50 was still alive then we would know that the time of death of the individual was in the interval $(50,\infty)$ but we would not know the exact age of death. Some names in common usage for various types of censoring are:

- Right-censoring: the censoring mechanism cuts short observation on a certain date so that censored individuals are known to have survived past that date.
- Left-censoring: data are left censored if we do not know when their condition started. For example if we were looking at survival past onset of HIV and we were tested at regular intervals, then individuals who were HIV positive at the time of the first test are left censored.
- Interval censoring: data are interval censored if we only know that their survival time falls into a certain interval. For example, in the previous HIV example, individuals who change HIV status between screens are interval censored.
- Random censoring: random censoring occurs when an individual is lost to a study for reasons completely unrelated to the disease under study. For example, in a study of the effect of a carcinogen on rats, a cage was lost when the water supply was lost over the weekend. Another example was an open door which resulted in escapes. In human studies, people can leave the area or die from unrelated causes, but care must be taken that the causes are completely unrelated.
- Informative and non-informative censoring: To continue the last example suppose we were looking at death from lung cancer. A death in a car accident where the lung cancer sufferer is not the driver might be an example of non-informative censoring since it is unlikely to have anything to do with the time that the patient might have ultimately died of lung cancer. However a death from

heart failure might be regarded as informative since lung stress places additional stress on the heart and can lead to heart failure. So death from heart failure can be regarded as informative.

- Type I censoring: when a study is terminated at a fixed predetermined time, it is called Type I censoring. Note that this is an example of right censoring. In the case of Type I censoring, the observed number of deaths is a random quantity.
- Type II censoring: when a study is continued until a fixed predetermined number of deaths have been observed, it is called Type II censoring. In this type of study the length of the trial is the random quantity, not the number of deaths.

Example 1.3 Acute myelogenous leukemia

A clinical trial was run to see if maintenance chemotherapy affected the progression of acute myelogenous leukemia. Patients in remission were assigned at random to a treatment group who received the maintenance chemotherapy and a control group who did not. The trial was Type I censored. The data comes from

Embury, SH, Elias, L, Heller, PH, Hood, CE, Greenberg, PL, and Schrier, SL (1977). *Remission maintenance therapy in acute myelogenous leukemia.* Western Journal of Medicine, 126:267-272.

The data is

```
> leukemia
   time status          group
 1    9      1     Maintained
 2   13      1     Maintained
 3   13      0     Maintained
 4   18      1     Maintained
 5   23      1     Maintained
 6   28      0     Maintained
 7   31      1     Maintained
 8   34      1     Maintained
 9   45      0     Maintained
10   48      1     Maintained
11  161      0     Maintained
12    5      1  Nonmaintained
13    5      1  Nonmaintained
14    8      1  Nonmaintained
15    8      1  Nonmaintained
16   12      1  Nonmaintained
17   16      0  Nonmaintained
18   23      1  Nonmaintained
```

19	27	1 Nonmaintained
20	30	1 Nonmaintained
21	33	1 Nonmaintained
22	43	1 Nonmaintained
23	45	1 Nonmaintained

The end point of interest in this example is relapse. The status variable is 1 if a relapse is observed and 0 if censored. Observations 1 through 11 are the treatment group and the others are the control group. If we use a * to denote a censored individual, then the data is

Table 1.1 Leukemia Data

Maintenance Chemotherapy	Control
9	5
13	5
13*	8
18	8
23	12
28*	16*
31	23
34	27
45*	30
48	33
161*	43
	45

The aim of the next section is to extend the empirical cdf estimator to censored data.

2. Kaplan-Meier Estimator

In this section we will discuss the generalization of the empirical cdf. For most of the section we will assume that times of death are not tied. If the sdf is continuous then there won't be any ties. Of course, in practice, real data will often have ties and we will indicate how to modify the formulae for ties. The other assumption that we will make is that if a censoring time is tied with a time of death, then the censored individual will be assumed to have been alive at the time of death. This is a reasonable assumption since the censored individual must have survived at least a small increment of time past the censoring. We will use the following notation.

2.1 Notation

The following notation will be used throughout this section:

- The total number of individuals in the study is denoted by N.
- The total number of deaths in the study is denoted by m.

- The k distinct times of deaths are $t_1 < t_2 < \ldots < t_k$. Note that if there are no tied times of death, then $k = m$.
- The number of deaths at time t_j is d_j, so $d_1 + d_2 + \ldots + d_k = m$.
- The number of individuals known to be alive at time t_j is denoted by r_j. So all individuals who haven't died or been censored by t_j are counted in r_j.

2.3 Heuristic derivation of the Kaplan-Meier Estimator

Suppose we have divided the time line into very small intervals. The intervals are so small that at most one event happens in each interval. The continuity assumption can be used to justify this assumption. Now we ask what would be a reasonable estimate of surviving a small interval given that you were alive at the start of the interval. Let the number of individuals known to be alive at the start of the interval be r and let d be the number of deaths in the interval. Note that our assumption implies that d is either 0 or 1. Then it is quite reasonable to estimate the probability of dying in the interval given that you are alive at the beginning of the interval or the hazard by d/r and the survival probability for the interval by $1-d/r$. Note that the estimated survival probabilities are 1 for intervals which do not contain a death and $(r-1)/r$ for intervals with a death. So we would estimate the probability of surviving a sequence of intervals given that the individual was alive at the beginning of the sequence by

$$\prod_{\text{intervals}\, j} \frac{r_j - d_j}{r_j}$$

Next, we let the intervals get infinitesimally small. What results is a function that only changes at times of death. So it is a step function with steps at times of death. The actual expression for the estimate of the survival distribution function is

$$\hat{S}_T(t) = \prod_{t_j \le t} \frac{r_j - d_j}{r_j} \qquad (2.1)$$

which is the Kaplan-Meier estimator. Although we have been quite loose in our derivation, a more formal derivation would proceed on approximately the same lines. The end result is that since we haven't made any assumptions concerning the underlying distribution, it is common to call the KM estimator the 'non-parametric estimator' of the sdf. Note that the estimator is 1 for times prior to the first death, that is for $t < t_1$.

If the last observed event is not a death, then $r_k > 1$ and $(r_k - 1)/r_k > 0$ and hence $\hat{S}_T(t_k) = s > 0$. So for all $t > t_k$, we have that $\hat{S}_T(t) = s > 0$. This means that our estimate of surviving to $t > t_k$ remains stuck at $s > 0$ and hence the estimated probability of surviving forever is s. This is a like a negative estimate of a variance – it can't be true. So the convention that is usually adopted is to set the Kaplan-Meier estimate to 0 past the last observation or sometimes past the last death. The expression (2.1) also holds for tied times of death.

Example 2.1 (Example 1.3 continued)
We consider again the leukemia example and construct the Kaplan-Meier estimators. Consider just the treatment group. Recall that the times of relapse or censoring are 9, 13, 13*, 18, 23, 28*, 31, 34, 45*, 48,161*. The calculations for the Kaplan-Meier estimator are

Table 2.1: Leukemia data

Death times t_j	r_j	$(r_j - d_j)/r_j$	$\hat{S}_T(t) = \prod_{t_l \le t_j} \frac{r_l - d_l}{r_l}$
9	11	$\frac{10}{11}$	$\frac{10}{11}$
13	10	$\frac{9}{10}$	$\frac{10}{11}\frac{9}{10}$
18	8	$\frac{7}{8}$	$\frac{10}{11}\frac{9}{10}\frac{7}{8}\frac{6}{7}$
23	7	$\frac{6}{7}$	$\frac{10}{11}\frac{9}{10}\frac{7}{8}\frac{6}{7}$
31	5	$\frac{4}{5}$	$\frac{10}{11}\frac{9}{10}\frac{7}{8}\frac{6}{7}\frac{4}{5}$
34	4	$\frac{3}{4}$	$\frac{10}{11}\frac{9}{10}\frac{7}{8}\frac{6}{7}\frac{4}{5}\frac{3}{4}$
48	2	$\frac{1}{2}$	$\frac{10}{11}\frac{9}{10}\frac{7}{8}\frac{6}{7}\frac{4}{5}\frac{3}{4}\frac{1}{2}$

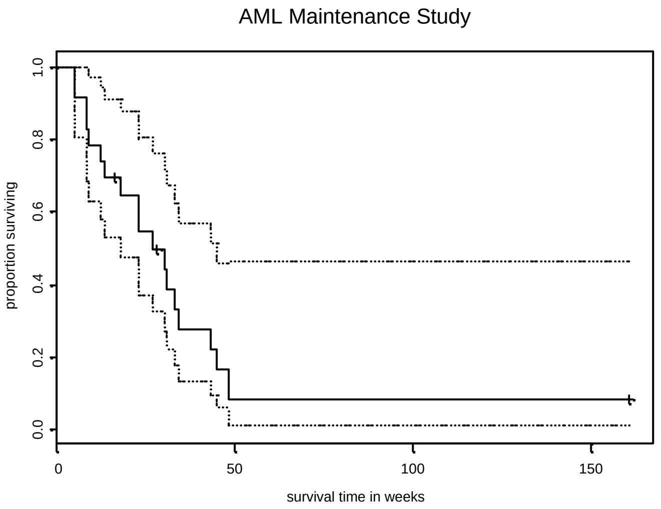
We can also use SPlus to calculate the KM estimator and look at the result.

First we will ignore the treatment groups and see what we get:

```
> fit_survfit(Surv(time,status),data=leukemia)
> summary(fit)
Call: survfit(formula = Surv(time, status), data = leukemia)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5     23       2   0.9130  0.0588       0.8049        1.000
    8     21       2   0.8261  0.0790       0.6848        0.996
```

```
  9    19     1  0.7826  0.0860     0.6310      0.971
 12    18     1  0.7391  0.0916     0.5798      0.942
 13    17     1  0.6957  0.0959     0.5309      0.912
 18    14     1  0.6460  0.1011     0.4753      0.878
 23    13     2  0.5466  0.1073     0.3721      0.803
 27    11     1  0.4969  0.1084     0.3240      0.762
 30     9     1  0.4417  0.1095     0.2717      0.718
 31     8     1  0.3865  0.1089     0.2225      0.671
 33     7     1  0.3313  0.1064     0.1765      0.622
 34     6     1  0.2761  0.1020     0.1338      0.569
 43     5     1  0.2208  0.0954     0.0947      0.515
 45     4     1  0.1656  0.0860     0.0598      0.458
 48     2     1  0.0828  0.0727     0.0148      0.462
> plot(fit,xlab="survival time in weeks",ylab="proportion
surviving", main="AML Maintenance Study")
```



Note the last time of death is 48 and the default in SPlus does not set the KM estimator to 0 past the last death. The other feature of the plot is that a pointwise 95% confidence interval for the estimate is given. At the moment, we don't know how to estimate the variance but we will find out shortly.

Now let's try taking account of the group structure:

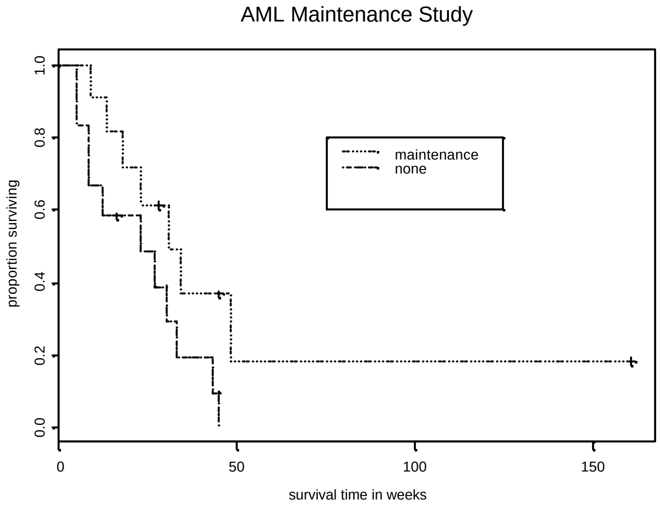
```
> fit2_survfit(Surv(time,status)~group,leukemia)
> summary(fit2)
Call: survfit(formula = Surv(time, status) ~ group, data =
leukemia)

              group=Maintained
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   9    11      1    0.909  0.0867     0.7541      1.000
  13    10      1    0.818  0.1163     0.6192      1.000
  18     8      1    0.716  0.1397     0.4884      1.000
```

```
   23     7     1   0.614  0.1526      0.3769       0.999
   31     5     1   0.491  0.1642      0.2549       0.946
   34     4     1   0.368  0.1627      0.1549       0.875
   48     2     1   0.184  0.1535      0.0359       0.944

              group=Nonmaintained
  time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5    12     2  0.8333  0.1076      0.6470       1.000
    8    10     2  0.6667  0.1361      0.4468       0.995
   12     8     1  0.5833  0.1423      0.3616       0.941
   23     6     1  0.4861  0.1481      0.2675       0.883
   27     5     1  0.3889  0.1470      0.1854       0.816
   30     4     1  0.2917  0.1387      0.1148       0.741
   33     3     1  0.1944  0.1219      0.0569       0.664
   43     2     1  0.0972  0.0919      0.0153       0.620
   45     1     1  0.0000      NA          NA          NA
```

```
> plot(fit2,xlab="survival time in weeks",ylab="proportion
surviving", main="AML Maintenance Study",lty=2:3)
> legend(c(75,125),c(.6,.8),c("maintenance","none"),lty=2:3)
```



2.4 d -method

We haven't explained where the estimate of the standard error of the KM estimator comes from. It is derived using the 'd -method' which is a way of approximating the mean and variance of a transformation of a random variable. Consider a function $U = g(Y)$ of a random variable Y which has mean m and variance s^2. We wish to approximate the mean and variance of U. Note that we are not assuming anything about the distribution of Y. A simple Taylor series expansion to the second term gives

$$g(Y) \approx g(m) + (Y - m)g'(m)$$

or

$$g(Y)-g(m)\approx(Y-m)g'(m).$$

So $E[g(Y)]\approx g(m)$ and $Var[g(Y)]\approx g'(m)^2 s^2$.

Example 2.2

Suppose Y has mean 5 and variance 10. Find the approximate mean and variance of $U=1/(1+Y^2)$. In this case, $g(y)=1/(1+y^2)$ and $g'(y)=-2y/(1+y^2)^2$. So $g'(5)=-2\times5/(1+5^2)^2=-10/676$ and $E[U]\approx1/26$, $Var(U)\approx(10/676)^2\times10$.

The accuracy of the approximation will depend on the linearity of the function near the mean of the random variable. In the multivariate version of the 'd -method', Y is vector of length p and with mean vector m and $p\times p$ covariance matrix Σ. The multivariate 'd -method' which is derived in the same manner using the first two terms of a Taylor is $E[g(Y)]\approx g(m)$ and $Var[g(Y)]\approx(\partial g/\partial m)^T\hat{\Sigma}\partial g/\partial m$, where $\partial g/\partial m$ is the vector of partial derivatives of g evaluated at the mean vector m.

Example 2.3

Suppose Y is a vector of length 2 with mean and variance

$$m=\begin{pmatrix}2\\4\end{pmatrix},\ \Sigma=\begin{pmatrix}2&3\\3&6\end{pmatrix}.$$

Find the approximate mean and variance of $U=1+Y_1/Y_2$, where $Y^T=(Y_1,Y_2)$. In this case $(\partial g/\partial m)^T=(1/m_2,-m_1/m_2^2)$. So the approximate mean of U is 1+2/4=1.5 and the approximate variance is

$$(1/4,-2/4^2)\begin{pmatrix}2&3\\3&6\end{pmatrix}\begin{pmatrix}1/4\\-2/4^2\end{pmatrix}=\frac{1}{32}.$$

The multivariate 'd -method' is applied to the KM estimator by writing it as

$$\hat{S}_T(t)=\prod_{t_j\le t}\hat{p}_j$$

where $\hat{p}_j$ is the estimate of the probability of surviving death j. We assume that the d_j's are independent binomials with parameters r_j and $\hat{p}_j$. Straightforward, but tedious, application of the multivariate ‘d -method’ which will be spelled out in more detail in the tutorial, leads to the following expression for an estimate of the variance of the KM estimator which is called Greenwood’s formula.

$$Var\left[\hat{S}_T(t)\right] \approx \hat{S}_T(t)^2 \sum_{t_j \le t} \frac{d_j}{r_j\left(r_j - d_j\right)},$$

which is the formula used in the SPlus output.

Example 2.4 Ovarian Cancer
A description of a trial to study ovarian cancer is given in the paper:

Edmunson, J. H., Fleming, T. R., Decker, D. G., Malkasian, G. D., Jefferies, J. A., Webb, M. J., and Kvols, L. K. (1979). *Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma vs. minimal residual disease.* Cancer Treatment Reports 63, 241-47.

The data from is the trial is available in SPlus in the object 'ovarian'. The variables in ovarian are

futime:	number of days in study
fustat:	indicator of death (1) or censoring (0)
age:	patient age in days/365.25
residual.dz:	an indicator of the extent of the residual disease
rx:	treatment given
ecog.ps:	a measure of performance

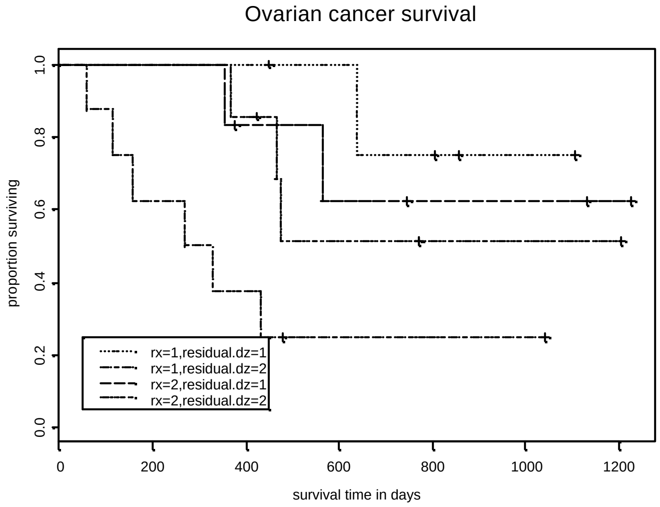
We could estimate four separate KM curves for this data, one for each possible combination of treatment (rx) and residual disease status (residual.dz).

```
> ovfit_survfit(Surv(futime,fustat)~rx+residual.dz,ovarian)
> ovfit
Call: survfit(formula = Surv(futime, fustat) ~ rx + residual.dz,
data = ovarian)

                     n events mean se(mean) median 0.95LCL 0.95UCL
rx=1, residual.dz=1 5      1  989      101     NA     638      NA
rx=1, residual.dz=2 8      6  430      131    298     156     NA
rx=2, residual.dz=1 6      2  943      161     NA     563      NA
rx=2, residual.dz=2 7      3  833      156     NA     464      NA
```

Note that the '+' in the formula results in a different model to ordinary linear regression. The survfit function interprets the operator '+' to mean fit all combinations of the terms so it is like the interaction operator. We can plot these KM's:

```
> plot(ovfit,xlab="survival time in days",
+ ylab="proportion surviving",
+ lty=2:5,
+ main="Ovarian cancer survival")
> legend(c(50,450),c(.05,.25),
+ c("rx=1,residual.dz=1","rx=1,residual.dz=2",
+ "rx=2,residual.dz=1","rx=2,residual.dz=2"),lty=2:5)
```



Confidence intervals for the survival distribution function at any time point are obtained in the usual as the estimate plus or minus the appropriate critical point of a standard normal by Greenwood's formula with the proviso that if either limit is greater than 1 or less than 0 it is changed to 1 or 0 to reflect the fact that we know that the sdf is between 0 and 1. So, subject to the previous comments concerning 0's and 1's, the $100(1-a)\%$ CI for $S_T(t)$ is

$$\hat{S}_T(t) \pm z_{a/2}\sqrt{\hat{S}_T(t)^2 \sum_{t_j \le t} \frac{d_j}{r_j(r_j - d_j)}}$$

The CI's for the ovarian data can be viewed in SPlus. Note that the default in SPlus produces intervals which are a transform of the intervals for the cumulative hazard. Specifically just for interest, the conf.type="log", which is the default, gives a confidence interval (C_L, C_U), where

$$C_L = \exp\left(-\log\left(\hat{S}_T(t)\right) - z_{a/2}\sqrt{\sum_{t_j \le t} \frac{d_j}{r_j(r_j - d_j)}} \right)$$

and

$$C_U = \min\left\{1, \exp\left(-\log\left(\hat{S}_T(t)\right) + z_{a/2}\sqrt{\sum_{t_j \le t} \frac{d_j}{r_j\left(r_j - d_j\right)}}\right)\right\}$$

To obtain intervals of the usual type you should specify conf.type="plain" in the call to survfit.

```
>ovfit_survfit(Surv(futime,fustat)~rx+residual.dz,
+ovarian,conf.type="plain")
> summary(ovfit)
Call: survfit(formula = Surv(futime, fustat) ~ rx + residual.dz,
data = ovarian, conf.type = "plain")

             rx=1, residual.dz=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  638    4       1     0.75   0.217       0.326           1

             rx=1, residual.dz=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   59     8       1   0.875   0.117      0.6458       1.000
  115     7       1   0.750   0.153      0.4499       1.000
  156     6       1   0.625   0.171      0.2895       0.960
  268     5       1   0.500   0.177      0.1535       0.846
  329     4       1   0.375   0.171      0.0395       0.710
  431     3       1   0.250   0.153      0.0000       0.550

             rx=2, residual.dz=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  353     6       1   0.833   0.152       0.535           1
  563     4       1   0.625   0.213       0.207           1

             rx=2, residual.dz=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  365     7       1   0.857   0.132       0.598       1.000
  464     5       1   0.686   0.186       0.321       1.000
  475     4       1   0.514   0.204       0.115       0.914
```

3. Nelson-Aalen (or Fleming-Harrington) Estimator

An alternative approach to estimating the survival distribution uses the relationship (1.1)

$$S_T(t) = \exp\left(-\int_0^t l_T(y)\,dy\right)$$

or $S_T(t) = \exp(-\Lambda_T(t))$. The cumulative hazard function is estimated by adding up the empirical hazard which has accumulated by time t which is

$$\hat{\Lambda}_T(t) = \sum_{t_j \le t} \frac{d_j}{r_j}. \qquad (3.1)$$

Plugging in this estimator, we obtain $\hat{S}_T(t) = \exp\left(-\hat{\Lambda}_T(t)\right)$. Variations on this estimator use different methods for handling ties. For example suppose two deaths were tied and the size of the risk set at that time was 10. The extra term in the sum would be 2/10. However if we could separate the two deaths, then the two extra terms would be 1/10+1/9 which is different to 2/10. The Fleming Harrington method always uses expressions of the form 1/10+1/9 which can be shown to lead to less bias. We will see in the tutorial that the KM and the FH estimators are approximately equal. We can also derive a variance estimate for the estimate of the cumulative hazard which is

$$Var\left(\hat{\Lambda}_T(t)\right) \approx \sum_{t_j \le t} \frac{d_j\left(r_j - d_j\right)}{r_j^3} \qquad (3.2)$$

The same comments as above would lead to the same approach as the Fleming-Harrington method for ties or

$$Var\left(\hat{\Lambda}_T(t)\right) \approx \sum_{t_j \le t} \frac{r_j - 1}{r_j^3} \qquad (3.3)$$

where we assume we have broken all ties. As we will see in the tutorial, using the 'd -method', it can be shown that

$$Var\left(\exp\left(-\hat{\Lambda}_T(t)\right)\right) \approx \exp\left(-2\hat{\Lambda}_T(t)\right) \sum_{t_j \le t} \frac{r_j - 1}{r_j^3}$$

or the variance of the Fleming-Harrington estimator is

$$Var\left[\hat{S}_T(t)\right] \approx \hat{S}_T(t)^2 \sum_{t_j \le t} \frac{r_j - 1}{r_j^3}.$$

SPlus can also calculate the Fleming-Harrington estimate of the survival. In Splus, the estimate of the integrated hazard with ties unbroken is obtained using the argument type="fleming-harrington". The estimator with the ties broken is obtained using type="fh2". The usual nomenclature in the literature is Nelson-Aaleen for the former and Fleming-Harrington for the latter but there is some confusion. So for the purposes of this course, to obtain what we will call the Nelson-Aaleen use type="fleming-harrington" and to obtain the Nelson-Aaleen use type="fh2". Another issue in Splus is that the quoted standard error for both of these estimators does not use

$$Var\left[\hat{S}_T(t)\right] \approx \hat{S}_T(t)^2 \sum_{t_j \le t} \frac{r_j - 1}{r_j^3}.$$

Instead it uses Greenwood's formula in for all 'types' as the default. The only alternative option is to use the Tsiatis formula (invoked by error="tsiatis"), which is yet another variation.

Example 3.1 More on the AML Study
The Nelson-Aaleen estimates for the leukemia data can be found as follows:

```
> summary(survfit(Surv(time,status)~group,
+ data=leukemia,type="fleming-harrington")
+ )
Call: survfit(formula = Surv(time, status) ~ group, data =
leukemia, type= "fleming-harrington")

            group=Maintained
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   9    11      1    0.913  0.0871       0.7575       1.000
  13    10      1    0.826  0.1174       0.6253       1.000
  18     8      1    0.729  0.1422       0.4974       1.000
  23     7      1    0.632  0.1572       0.3882       1.000
  31     5      1    0.517  0.1731       0.2687       0.997
  34     4      1    0.403  0.1781       0.1695       0.958
  48     2      1    0.244  0.2038       0.0477       1.000

            group=Nonmaintained
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   5    12      2   0.8465   0.109       0.6572       1.000
   8    10      2   0.6930   0.141       0.4645       1.000
  12     8      1   0.6116   0.149       0.3791       0.987
  23     6      1   0.5177   0.158       0.2849       0.941
  27     5      1   0.4239   0.160       0.2021       0.889
  30     4      1   0.3301   0.157       0.1300       0.838
  33     3      1   0.2365   0.148       0.0692       0.808
  43     2      1   0.1435   0.136       0.0225       0.914
  45     1      1   0.0528     Inf       0.0000       1.000
```

Recall the KM estimates:

```
> summary(survfit(Surv(time,status)~group,
+ data=leukemia))
```

```
Call: survfit(formula = Surv(time, status) ~ group, data =
leukemia)

          group=Maintained
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   9     11       1    0.909  0.0867      0.7541       1.000
  13     10       1    0.818  0.1163      0.6192       1.000
  18      8       1    0.716  0.1397      0.4884       1.000
  23      7       1    0.614  0.1526      0.3769       0.999
  31      5       1    0.491  0.1642      0.2549       0.946
  34      4       1    0.368  0.1627      0.1549       0.875
  48      2       1    0.184  0.1535      0.0359       0.944

          group=Nonmaintained
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   5     12       2   0.8333  0.1076      0.6470       1.000
   8     10       2   0.6667  0.1361      0.4468       0.995
  12      8       1   0.5833  0.1423      0.3616       0.941
  23      6       1   0.4861  0.1481      0.2675       0.883
  27      5       1   0.3889  0.1470      0.1854       0.816
  30      4       1   0.2917  0.1387      0.1148       0.741
  33      3       1   0.1944  0.1219      0.0569       0.664
  43      2       1   0.0972  0.0919      0.0153       0.620
  45      1       1   0.0000      NA          NA          NA
```

By inspection, we see that differences between the two estimators are small.

4. The Cox Proportional Hazards Regression Model

4.1 Introduction

Often we are interested in the effect of covariates on the survival distribution. Covariates are variables other than the survival time measured on individuals which we believe may influence the survival prospects of individuals. We have seen how to estimate separate survival curves for each level of a discrete covariate in the last section. This will not be possible for continuous covariates. In any case, we still do not know how to compare survival curves. In many situations, the issue of interest is not the underlying survival curve, but the effect of covariates. For instance, recall the ovarian cancer example. We might be interested in the effect of the age of the woman at diagnosis on the survival time. Age is a continuous random variable, so if we wished to adopt the approach in the previous section, then we would need to stratify according to some age grouping and then compute the KM for each age strata. Even then we would have no formal way of assessing the difference between the strata with regard to survival.

Essentially we are posing a regression problem. We have 'independent' variables which influence the distribution of the 'dependent' variable and we wish to assess that effect. We are used to formulating regression models for normal data and more generally GLM's. It turns out that parametric regression modeling for survival data with censoring is very challenging. Fortunately there is a much more tractable alternative method. The method avoids specifying a particular family of distributions and instead concentrates on estimating the effect of the covariates.

4.2 Proportional hazards assumption

The key assumption is called the 'Proportional Hazards Assumption'. Suppose that the survival time of an individual is denoted by T and the covariates by z which is a p dimensional vector. Then the 'Proportional Hazards Assumption' is that the hazard is

$$l_T(t;z) = l_0(t)\exp\left(b^T z\right), \qquad (4.1)$$

where b is a p dimensional vector of unknown parameters. The function $l_0(t)$ is called the base hazard and is the hazard function of an individual with covariate $z = 0$. The key feature of (4.1) is that the effect of the covariates is to multiply the underlying hazard. If the corresponding survival distribution is $S_T(t;z)$, then

$$\begin{aligned} S_T(t;z) &= \exp\left(-\int_0^t l_T(y;z)dy \right) \\ &= \exp\left(-\int_0^t l_0(y)\exp\left(b^T z\right)dy \right) \\ &= \exp\left(-\exp\left(b^T z\right)\int_0^t l_0(y)dy \right) \\ &= \left(\exp\left(-\int_0^t l_0(y)dy \right)\right)^{\exp\left(b^T z\right)} \\ &= S_0(t)^{\exp\left(b^T z\right)} \end{aligned}$$

where

$$S_0(t) = \exp\left(-\int_0^t l_0(y)dy \right)$$

is the underlying survival distribution function. So the effect of the covariates is to raise the underlying survival distribution to a power. Note that this means that survival distributions for different values of the covariates can never cross – survival is either always better or always worse. For example if the covariate is type of treatment, then you are either always better or always worse than under another treatment. This rules out the possibility that an aggressive treatment may have a high initial fatality rate but much better long term survival. This problem can be fixed to some extent by extending the model to allow time dependent covariates.

4.3 Partial likelihood

In this last section we introduced the proportional hazards model. However we have not discussed why the model has been so widely applied. The reason is that it leads to a 'non-parametric' or 'semi-parametric' method of estimating b which avoids specification and even estimation of $l_0(t)$. The key to the method is the observation that if we take the ratio of the hazards at time t for two individuals with covariates z_1 and z_2 respectively,

$$\begin{aligned}\frac{l(t;z_1)}{l(t;z_2)} &= \frac{l_0(t)\exp\left(b^T z_1\right)}{l_0(t)\exp\left(b^T z_2\right)} \\ &= \frac{\exp\left(b^T z_1\right)}{\exp\left(b^T z_2\right)}\end{aligned}$$

which does not involve $l_0(t)$. So if we could find a likelihood which only involves ratios of hazard functions then we would avoid consideration of $l_0(t)$. The solution is to focus on a time of death, $t_{(j)}$ say where we use (j) as a subscript to denote the individual who actually died at $t_{(j)}$, and ask:

> "Among all individuals who were under observation at time $t_{(j)}$, what is the probability that individual (j) died."

So what we are actually asking is what is the conditional probability that individual (j) dies at $t_{(j)}$ given the fact that there is a single death at $t_{(j)}$ and the set of individuals under observation or 'at risk' at $t_{(j)}$. We denote the set of individuals 'at risk' at time t by $\Re(t)$. Now the probability that an individual with covariate z fails in the interval $(t, t+u)$ for small u is approximately $l(t;z)u$ and the probability that there is more than one death in $(t, t+u)$ will be of order u^2 which will be very small. So

$$
\begin{aligned}
P\left[\begin{array}{l}\text{Individual } (j) \text{ dies in } \left(t_{(j)}, t_{(j)}+u\right) \\ \left|\Re\left(t_{(j)}\right) \text{and a death in } \left(t_{(j)}, t_{(j)}+u\right)\right.\end{array}\right] &\approx \frac{l\left(t_{(j)}; z_{(j)}\right)u}{\sum_{l \in \Re\left(t_{(j)}\right)} l\left(t; z_l\right)u} \\
&= \frac{l_0\left(t_{(j)}\right)\exp\left(b^T z_{(j)}\right)}{\sum_{l \in \Re\left(t_{(j)}\right)} l_0\left(t_{(j)}\right)\exp\left(b^T z_l\right)} \\
&= \frac{\exp\left(b^T z_{(j)}\right)}{\sum_{l \in \Re\left(t_{(j)}\right)} \exp\left(b^T z_l\right)} \\
&= p_{(j)}(b).
\end{aligned}
\tag{4.2}
$$

The thing to notice about this expression is that both u and $l_0(t)$ do not appear in the final line. So a 'likelihood' based on conditional probabilities of the sort above will not involve $l_0(t)$ and estimation of b can proceed without specifying the underlying family. Recall that the likelihood of a random sample is the product of the densities of the individual observations. Here we will use the product of the terms (4.2) over all times of death. This is not a likelihood in the usual sense, but has been dubbed a 'partial likelihood'. A lot of work has been done showing that the partial likelihood can be treated in the same way as a likelihood and the usual properties hold. So the partial likelihood for b is

$$
PL(b) = \prod_{j=1}^{k} \frac{\exp\left(b^T z_{(j)}\right)}{\sum_{l \in \Re\left(t_{(j)}\right)} \exp\left(b^T z_l\right)} \tag{4.3}
$$

The maximum likelihood estimate of an unknown parameter is the value of the parameter that maximizes the likelihood. By analogy the estimate of b from the partial likelihood will be the value of b which maximizes the partial likelihood. The most common way of handling ties is using Breslow's approximation. (Note that the default in SPlus is not Breslow's approximation. SPlus uses the Efron method which is exact but much more demanding computationally. See the SPlus documentation for more information. In particular to use the Breslow approximation, add the argument 'method="breslow" to the call to coxph().) Suppose the z 's for the d_j deaths at $t_{(j)}$ are $z_{(j),1}, z_{(j),2}, \ldots, z_{(j),d_j}$ and let $s_{(j)} = z_{(j),1} + z_{(j),2} + \ldots + z_{(j),d_j}$ be the sum of the covariates of the individuals that died at $t_{(j)}$. Then the Breslow approximation to the partial likelihood for tied deaths is

$$PL(b) = \prod_{j=1}^{k} \frac{\exp\left(b^T s_{(j)}\right)}{\left(\sum_{l \in \Re\left(t_{(j)}\right)} \exp\left(b^T z_l\right)\right)^{d_j}}. \qquad (4.4)$$

Example 4.1
A study was conducted to look at the effect of a treatment on the survival time of liver transplant patients. The patients weight was also thought to influence survival. The data was

Patient #	Survival Time	Patient weight	Treatment	(z_1, z_2)	$(b_1, b_2)\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$
1	3	83	no	(0,83)	$83b_2$
2	6*	58	yes	(1,58)	$b_1 + 58b_2$
3	9	68	no	(0,68)	$68b_2$
4	11	73	yes	(1,73)	$b_1 + 73b_2$
5	14	75	no	(0,75)	$75b_2$
6	14	68	yes	(1,68)	$b_1 + 68b_2$
7	14*	49	yes	(1,49)	$b_1 + 49b_2$
8	16	86	no	(0,86)	$86b_2$

The distinct times of death are 3, 9, 11, 14, 16. Note that there are tied deaths at 14 so we will use equation (4.4) when constructing the partial likelihood. The censoring at 14 is assumed to take place after the deaths. The covariates are z_1 which is 1 for an individual with the treatment and 0 otherwise and z_2 which is the weight of the patient. If the vector of unknown parameters is b where $b^T = (b_1, b_2)$, then the terms in the partial likelihood are:

$t_{(j)}$	Patients in $\Re\left(t_{(j)}\right)$	d_j	$s_{(j)}$	$b^T s_{(j)}$
3	1 to 8	1	(0,83)	$83b_2$
9	3 to 8	1	(0,68)	$68b_2$
11	4 to 8	1	(1,73)	$b_1 + 73b_2$
14	5 to 8	2	(1,143)	$b_1 + 143b_2$
16	8	1	(0,86)	$86b_2$

Table continued:

$\dfrac{\exp\left(b^T s_{(j)}\right)}{\left(\sum_{l\in\Re\left(t_{(j)}\right)}\exp\left(b^T z_l\right)\right)^{d_j}}$
$\dfrac{e^{83b_2}}{e^{83b_2}+e^{b_1+58b_2}+e^{68b_2}+e^{b_1+73b_2}+e^{75b_2}+e^{b_1+68b_2}+e^{b_1+49b_2}+e^{86b_2}}$
$\dfrac{e^{68b_2}}{e^{68b_2}+e^{b_1+73b_2}+e^{75b_2}+e^{b_1+68b_2}+e^{b_1+49b_2}+e^{86b_2}}$
$\dfrac{e^{b_1+73b_2}}{e^{b_1+73b_2}+e^{75b_2}+e^{b_1+68b_2}+e^{b_1+49b_2}+e^{86b_2}}$
$\dfrac{e^{b_1+143b_2}}{\left(e^{75b_2}+e^{b_1+68b_2}+e^{b_1+49b_2}+e^{86b_2}\right)^2}$
$\dfrac{e^{86b_2}}{e^{86b_2}}$

The partial likelihood is the product of the 5 terms above.

Now that we have defined the partial likelihood and assumed that it has the same properties as the usual likelihood, we are still left with the problem of how to find $\hat{b}$ which is the value of the vector b which maximizes the partial likelihood $PL(b)$. Often in maximum likelihood estimation, the MLE must be found by numerical techniques such as Newton-Rhapson. It is almost always the case that $\hat{b}$ from the partial likelihood must be found by numerical techniques. Fortunately, these days, almost all statistical software packages will fit Cox Proportional Hazards Models and provide the estimate of $\hat{b}$.

Example 4.2 (4.1 continued)
Consider using SPlus to fit the proportional hazards model from the previous example. The relevant SPlus function is coxph(). All the syntax is the same as when using survfit().

```
> livertime_c(3,6,9,11,14,14,14,16)
> livertime
[1]  3  6  9 11 14 14 14 16
> liverstatus_c(1,0,1,1,1,1,0,1)
> livertreat_c("No","Yes","No","Yes","No","Yes","Yes","No")
> weight_c(83,58,68,73,75,68,49,86)
> liver_data.frame(livertime,liverstatus,livertreat,weight)
> liver
  livertime liverstatus livertreat weight
```

```
1      3      1      No    83
2      6      0     Yes    58
3      9      1      No    68
4     11      1     Yes    73
5     14      1      No    75
6     14      1     Yes    68
7     14      0     Yes    49
8     16      1      No    86

> liverfit_coxph(Surv(livertime,liverstatus)~
+livertreat+weight,liver)

> liverfit
Call:
coxph(formula = Surv(livertime, liverstatus) ~ livertreat +
weight, data = liver)


               coef exp(coef) se(coef)       z    p
livertreat -0.0402     0.961   0.6801 -0.0591 0.95
    weight  0.0223     1.023   0.0567  0.3927 0.69

Likelihood ratio test=0.45  on 2 df, p=0.799  n= 8

> summary(liverfit)
Call:
coxph(formula = Surv(livertime, liverstatus) ~ livertreat +
weight, data = liver)

  n= 8

               coef exp(coef) se(coef)       z    p
livertreat -0.0402     0.961   0.6801 -0.0591 0.95
    weight  0.0223     1.023   0.0567  0.3927 0.69

           exp(coef) exp(-coef) lower .95 upper .95
livertreat     0.961      1.041     0.253      3.64
    weight     1.023      0.978     0.915      1.14

Rsquare= 0.055   (max possible= 0.863 )
Likelihood ratio test= 0.45  on 2 df,   p=0.799
Wald test            = 0.42  on 2 df,   p=0.81
Score (logrank) test = 0.43  on 2 df,   p=0.806

> summary(survfit(liverfit))
Call: survfit.coxph(object = liverfit)

 time n.risk n.event sur std.err lower 95% CI upper 95% CI
    3      8       1   0.887   0.107       0.6995        1.000
    9      6       1   0.755   0.154       0.5063        1.000
```

11	5	1	0.622	0.177	0.3561	1.000
14	4	2	0.356	0.180	0.1318	0.961
16	1	1	0.182	0.181	0.0257	1.000

There is a lot of information in this printout. The parameter estimates are $\left(\hat{b}_1, \hat{b}_2\right) = (-.0402,.0223)$. How should we interpret these estimates? Firstly, consider $\hat{b}_1$. This is the estimated coefficient of z_1 which is either 0 or 1 according to whether or not the patient received the treatment. The effect is to multiply the hazard by $e^{-.0402}$ or .961 for individuals who get the treatment. So treatment lowers the hazard by 3.9% - not a great amount. Since b_2 is the coefficient of weight in the model, an increase in weight of 1 kilogram multiplies the hazard by $e^{.0223}$ or 1.023 or 2.3%.

In the SPlus printout, there are various standard errors and test statistics quoted. It is beyond the scope of this course to give a complete derivation of the underlying theory. However we can give some indication of how the results arise. Recall in maximum likelihood in what was called a regular situation, the MLE was found by differentiating the log likelihood and equating it to zero. One way of approximating the distribution of the MLE was to use asymptotic theory which said that the MLE is approximately unbiased and normal with variance the inverse of the observed information matrix. The same result applies to the partial likelihood estimator. The estimator $\hat{b}$ is the solution of the efficient score equation, $u(b) = 0$, where the p dimensional efficient score is

$$u(b) = \frac{\partial \log PL(b)}{\partial b}$$

$$= \begin{pmatrix} \dfrac{\partial \log PL(b)}{\partial b_1} \\ \dfrac{\partial \log PL(b)}{\partial b_2} \\ \vdots \\ \dfrac{\partial \log PL(b)}{\partial b_p} \end{pmatrix}$$

In the case of no ties, after some algebra, it can be shown that the efficient score statistic is

$$u(b) = \sum_{j=1}^{k} \left(z_{(j)} - m_{(j)}(b)\right)$$

where $m_{(j)}(b)$ is the conditional expectation of the covariates of the individual who dies at $t_{(j)}$,

$$m_{(j)}(b) = \sum_{l \in \Re\left(t_{(j)}\right)} p_{(j)l}(b) z_l .$$

Note that $p_{(j)l}(b)$ is the conditional probability that individual l dies at $t_{(j)}$ given $\Re\left(t_{(j)}\right)$ and a death at $t_{(j)}$, or the term used in the Partial Likelihood,

$$p_{(j)l}(b) = \frac{\exp\left(b^T z_l\right)}{\sum_{n \in \Re\left(t_{(j)}\right)} \exp\left(b^T z_n\right)} .$$

Note that in the case $b = 0$, $p_{(j)l} = p_{(j)l}(0) = 1/r_j$ where r_j is the number in $\Re\left(t_{(j)}\right)$ and that $m_{(j)}$ is the simple average of the covariates of the individuals in $\Re\left(t_{(j)}\right)$.

The sample information matrix is the $p \times p$ matrix $I\left(\hat{b}\right)$of second partial derivatives, which has ij^{th} entry

$$-\frac{\partial^2 \log PL(b)}{\partial b_i \partial b_j}$$

evaluated at $\hat{b}$. In the case of no ties, after some algebra, it can be shown that the sample information matrix is

$$I(b) = \sum_{j=1}^{k} V_{(j)}(b),$$

where $V_{(j)}(b)$ is the conditional variance of the covariates of the individual who dies at $t_{(j)}$,

$$V_{(j)}(b) = \sum_{l \in \Re\left(t_{(j)}\right)} p_{(j)l}(b)\left(z_l - m_{(j)}(b)\right)\left(z_l - m_{(j)}(b)\right)^T .$$

Inference can be conducted in the usual manner using the efficient score statistic and the sample information matrix. The score statistic can be used in score tests which is sometimes called the log rank test in this application. If the estimator $\hat{b}$ is used in a quadratic form in $I(b)$, then the Wald test statistic results. The partial likelihood can also be used in likelihood ratio tests. Recall that if L_p is the likelihood evaluated at the maximum likelihood estimate when there are p parameters, then to check the significance of adding an additional q terms to the model, we can use the Likelihood Ratio Test Statistic $-2\log\left(L_p/L_{p+q}\right)$ which has an approximate c_q^2 distribution when the additional parameters are all 0.

So far in the discussion we have gone to great lengths to avoid estimating $l_0(t)$. Of course it is often of great interest to estimate $l_0(t)$ and then to plug in $\hat{l}_0(t)$ and $\hat{b}$ into the relationship

$$\begin{aligned} S_T(t;z) &= \left(\exp\left(-\int_0^t l_0(y)dy\right)\right)^{\exp\left(b^T z\right)} \\ &= S_0(t)^{\exp\left(b^T z\right)} \end{aligned}$$

to obtain an estimated survival curve for covariate value z. In the Kaplan-Meier situation, we estimated the probability of surviving death (j) by $(r_j - 1)/r_j$. It turns out that an appropriate estimator for an individual with covariates $z = 0$ is

$$\left(1-\frac{\exp\left(\hat{b}^T z_{(j)}\right)}{\sum_{l\in\Re\left(t_{(j)}\right)}\exp\left(\hat{b}^T z_l\right)}\right)^{\exp\left(-\hat{b}^T z_{(j)}\right)}$$

and the resulting estimate of $S_0(t)$ is

$$\hat{S}_0(t) = \prod_{t_{(j)}\le t}\left(1-\frac{\exp\left(\hat{b}^T z_{(j)}\right)}{\sum_{l\in\Re\left(t_{(j)}\right)}\exp\left(\hat{b}^T z_l\right)}\right)^{\exp\left(-\hat{b}^T z_{(j)}\right)}.$$

The estimated survival curve for covariate value z is then

$$\hat{S}_T(t;z) = \hat{S}_0(t)^{\exp\left(\hat{b}^T z\right)}.$$

Example 4.3 (Example 4.2 continued)
The SPlus output from the previous example containing the following information concerning test statistics:

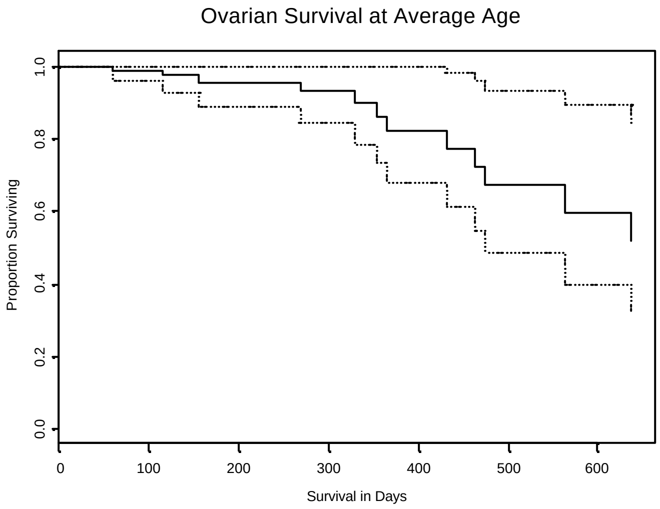
```
Likelihood ratio test= 0.45  on 2 df,   p=0.799
Wald test            = 0.42  on 2 df,   p=0.81
Score (logrank) test = 0.43  on 2 df,   p=0.806
```

All of these tests are tests of the hypothesis $H_0 : b_1 = b_2 = 0$ versus $H_1 : b_1 \neq 0 \vee b_2 \neq 0$. Note that the test statistics are almost identical – this is usually the case. In this case, we would conclude that neither treatment or weight is having a significant impact on survival.

Example 4.4 Ovarian cancer example continued
Earlier we fitted a range of KM estimates to the ovarian cancer data. Now let's try fitting a proportional hazards model.

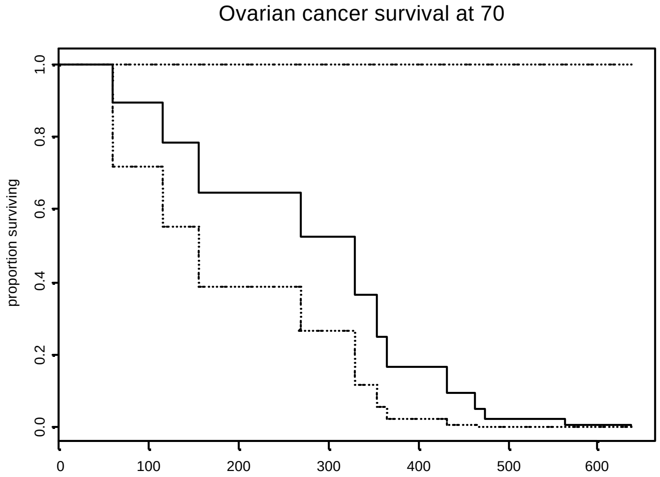
```
> summary(ovarian)
     futime          fustat          age        residual.dz
   Min.:  59.0     Min.:0.0000     Min.:38.89     Min.:1.000
 1st Qu.: 368.0   1st Qu.:0.0000  1st Qu.:50.17   1st Qu.:1.000
  Median: 476.0    Median:0.0000   Median:56.85
Median:2.000
   Mean: 599.5     Mean:0.4615     Mean:56.17     Mean:1.577
 3rd Qu.: 794.8   3rd Qu.:1.0000  3rd Qu.:62.38   3rd Qu.:2.000
   Max.:1227.0     Max.:1.0000     Max.:74.50     Max.:2.000
       rx        ecog.ps
   Min.:1.0     Min.:1.000
 1st Qu.:1.0   1st Qu.:1.000
  Median:1.5    Median:1.000
   Mean:1.5      Mean:1.462
 3rd Qu.:2.0   3rd Qu.:2.000
   Max.:2.0      Max.:2.000
> ovfit1_coxph(Surv(futime,fustat)~age,ovarian)
> summary(survfit(ovfit1))
Call: survfit.coxph(object = ovfit1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   59     26       1    0.988  0.0142        0.961        1.000
  115     25       1    0.974  0.0244        0.927        1.000
  156     24       1    0.955  0.0364        0.886        1.000
  268     23       1    0.933  0.0482        0.844        1.000
  329     22       1    0.897  0.0621        0.783        1.000
  353     21       1    0.862  0.0724        0.732        1.000
  365     20       1    0.824  0.0819        0.678        1.000
  431     17       1    0.775  0.0934        0.612        0.982
  464     15       1    0.724  0.1032        0.548        0.958
  475     14       1    0.673  0.1112        0.487        0.931
  563     12       1    0.596  0.1226        0.398        0.892
  638     11       1    0.520  0.1287        0.321        0.845
> plot(survfit(ovfit1),xlab="Survival in Days",
+ ylab="Proportion Surviving",main="Ovarian Survival at
Average Age")
```



```
# Note that the plot is at the average value of the covariates or
the average age. To plot the estimate of the survival distribution
at a particular age we need to supply a new dataframe with the
same column names as ovarian with the desired covariate values

> temp_ovarian
> temp_ovarian[1,]
> temp
  futime fustat     age residual.dz rx ecog.ps
1     59      1 72.3315           2  1       1
> temp[1,3]_70
> temp
  futime fustat age residual.dz rx ecog.ps
1     59      1  70           2  1       1
> mean(ovarian[,"age"])
[1] 56.16544
>  plot(survfit(ovfit1,newdata=temp),xalb="survival in days",
+ ylab="proportion surviving",main="Ovarian cancer survival
at 70")
```



```
>ovfit2_coxph(Surv(futime,fustat)~age+residual.dz+rx+ecog.ps
+ ,ovarian)
> summary(ovfit2)
Call:
coxph(formula = Surv(futime, fustat) ~ age + residual.dz  + rx +
ecog.ps, data = ovarian)

n= 26

              coef exp(coef) se(coef)      z      p
        age  0.125     1.133   0.0469  2.662 0.0078
residual.dz  0.826     2.285   0.7896  1.046 0.3000
         rx -0.914     0.401   0.6533 -1.400 0.1600
    ecog.ps  0.336     1.400   0.6439  0.522 0.6000

            exp(coef) exp(-coef) lower .95 upper .95
        age     1.133      0.883     1.033      1.24
residual.dz     2.285      0.438     0.486     10.74
         rx     0.401      2.496     0.111      1.44
    ecog.ps     1.400      0.714     0.396      4.94

Rsquare= 0.481   (max possible= 0.932 )
Likelihood ratio test= 17  on 4 df,   p=0.0019
Wald test            = 14.2  on 4 df,   p=0.00654
Score (logrank) test = 20.8  on 4 df,   p=0.000345

#We can construct a likelihood ratio test of the full model vs the
reduced model

> -2*(ovfit1$loglik[2]-ovfit2$loglik[2])
[1] 2.749708
> pchisq(2.749708,df=3)
[1] 0.5681541
```

AUSTRALIAN NATIONAL UNIVERSITY

SCHOOL OF FINANCE AND APPLIED STATISTICS

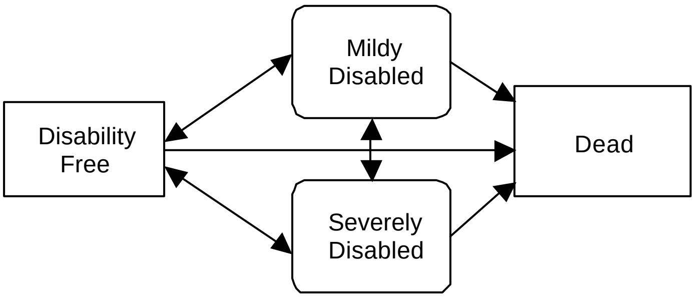
SURVIVAL MODELS 1 (STAT3032/STAT8042) BIOSTATISTICS (STAT8003) SURVIVAL MODELS 1 (STAT3032/STAT8042) BIOSTATISTICS (STAT8003)

Part 2

Sections 5 to 8

5. Two state Markov models

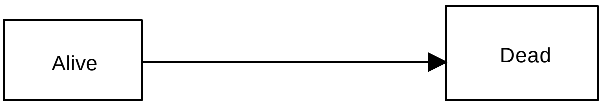
The aim of the next two sections is to develop multistate Markov models for survival and health data. The underlying model is that an individual can be in and move between various states. For example, recently there has been great interest in the amount of time that an individual will spend in a disability free state as compared to mildly disabled or severely disabled or eventually dead. We could depict the states in the following diagram:



The aim is to estimate various aspects of the probabilities associated with this diagram using data on the amount of time that individuals spend in the states and the transitions between the states. The models are called 'multistate' for obvious reasons. The 'markov' assumption is that the probability of moving to a new state depends only on the current state and not the previous history.

In this section, a different approach is taken to estimating survival. Specific forms for hazard functions are postulated and the maximum likelihood estimates are found. This type of analysis is often done on stratified data where we have grouped the data into homogeneous groups. So within the groups, the individuals are sufficiently similar (or the covariates are sufficiently close) so that it can be assumed that the hazard function is the same for all individuals. The alternative is the method from the last section where we allow the hazard function to vary across all individuals according as the covariates.

Initially in this section we will consider a two state Markov model consisting of only alive or dead. Once we have settled the concepts and estimation procedures we will move on to true multistate models. The simplified graphic for the two state model is:



The theory in this section will focus on a particular age, x say, and we are interested in estimating the hazard function $l(y)$ for $x \le y < x+1$. We will suppose that we observe a number of independent individuals, all with the hazard function $l(y)$, for some subset of the age interval $[x, x+1)$. Consider one individual. We would like to derive the likelihood (or pdf) for the observed results from a single individual over the age interval $[x, x+1)$. Let $x+a$ be the age when the individual comes under observation and $x+b$ the age at which the individual would stop being observed or would be Type I censored (in that age group). So a will be 0 if the individual is already under observation at the time they turn age x and will be between 0 and 1 if they commence observation in the age group sometime after they turn x. Also if individuals were not due for censoring until the end of the age group, then b will be 1 and otherwise it will be less than 1 but greater than a. We will typically make the assumption that if individuals are random censored in the age interval, then they would have survived until the end of the age interval. Note that if the individual dies before $x+b$, then the individual will not be monitored past death.

Suppose V is time from commencement of observation to observed death in $[x, x+1)$ or $x+b$. Let d be an indicator variable which is 1 if the individual is observed to die in $[x, x+1)$ and 0 otherwise. In order to use likelihood theory, we need to derive the likelihood or pdf of (V, d). The pdf has two components:

- The individual is observed to survive from age $x+a$ to age $x+a+v$. If we use the usual notation that ${}_t p_x$ is the probability of surviving to age $x+t$ given survival to x, then the contribution to the likelihood is ${}_v p_{x+a}$.
- If d is 1, then the individual dies at $x+a+v$. Since we have already taken into account survival until $x+a+v$, the appropriate contribution to the likelihood is the hazard $l\left(x+a+v\right)^d$. Note that the effect of raising the hazard to the power d is to make the contribution 1 if the individual doesn't die.

So the individual's contribution to the likelihood is ${}_v p_{x+a}\, l\left(x+a+v\right)^d$. Once we multiply across all individuals observed in the age group, we have the likelihood for l over the age interval $\left[x, x+1\right)$. To see that it is a likelihood for l, we need to express ${}_v p_{x+a}$ as a function of l. The same method is used to derive the relationship (1.1) between the sdf and the hazard function,

$$S_T\left(t\right) = \exp\left(-\int_0^t m_T\left(y\right)dy \right)$$

can be used to show that

$${}_v p_{x+a} = \exp\left(-\int_{x+a}^{x+a+v} l\left(y\right)dy \right).$$

So the contribution from the individual to the likelihood for l over the age interval $\left[x, x+1\right)$ is

$$L = \exp\left(-\int_{x+a}^{x+a+v} l\left(y\right)dy \right) l\left(x+a+v\right)^d \qquad (5.1)$$

Then next issue to consider is how to use the likelihood term (5.1). We can either use a 'semi-parametric' approach or we can assume a specific form for l. The approach taken for the remainder of this section is to assume that the hazard function is constant, l_x say, between age x and $x+1$ where x is the age currently under study. Note that this means that we are assuming that the survival random variable is piecewise exponential. We now wish to find the MLE of l_x. The likelihood term (5.1) becomes

$$\begin{aligned} L &\propto \exp\left(- \int_{x+a}^{x+a+v} l_x dy \right) l_x^{\,d} \\ &= \exp(- l_x v) l_x^{\,d}. \end{aligned} \qquad (5.2)$$

Multiplying the likelihood terms across all individuals who are observed in $[x, x+1)$ we obtain the likelihood

$$L \propto \exp(- l_x v_x) l_x^{\,d_x} \qquad (5.3)$$

where v_x is the total of v across individuals, sometimes called the total time on test, and d_x is the total number of deaths in the age interval. Note that (5.3) has the form of an exponential likelihood. In the usual way, by taking logs, differentiating and equating to 0, we obtain that the MLE of l_x is

$$\hat{l}_x = \frac{d_x}{v_x}.$$

In the presence of censoring, the distribution of $\hat{l}_x$ is very difficult to determine, however we can find an approximate distribution using the Central Limit Theorem. Consider the likelihood from a single individual (5.3). Then

$$P(d = 1) = \int_0^{b-a} l \exp(- l v) dv$$

and $P(d = 0) = P(v \geq b - a) = \exp(- l(b - a))$. So

$$\int_0^{b-a} l \exp(- l v) dv + \exp(- l(b - a)) = 1, \qquad (5.4)$$

which is of course obviously true. When (5.4) is differentiated with respect to l and the result multiplied by l, we obtain

$$\int_0^{b-a} l e^{-lv} dv - l \left\{ \int_0^{b-a} v l e^{-lv} dv + (b - a) e^{-l(b-a)} \right\} = 0. \qquad (5.5)$$

But

$$\begin{aligned} E[d] &= P(v < b - a) \\ &= \int_0^{b-a} l \exp(- l v) dv \end{aligned}$$

and

$$E[v] = \int_0^{b-a} v\, l \exp(-l\, v) dv + (b-a)\exp(-l\,(b-a)).$$

So $E[d - l\,v] = 0$ or $E[d] = l\,E[v]$. Dividing (5.5) by l, and letting $c = b - a$, we have

$$\int_0^c e^{-l\,v} dv - \int_0^c v\,l\,e^{-l\,v} dv - ce^{-l\,c} = 0 .$$

Differentiating this with respect to l, we obtain

$$-2\int_0^c ve^{-lv} dv + \int_0^c v^2\,l\,e^{-lv} dv + c^2 e^{-lc} = 0 ,$$

or

$$E[v^2] - 2E[dv]/l = 0 .$$

So since $E[d^2] = E[d]$, it follows that

$$E[(lv)^2] - 2E[d\,lv] + E[d^2] = E[d],$$

and since $E[d - l\,v] = 0$,

$$\begin{aligned} Var[d - l\,v] &= E[(d - lv)^2] \\ &= E[d]. \end{aligned}$$

So adding across independent individuals, we obtain

$$E[d_x - l_x v_x] = 0 ,\ Var[d_x - l_x v_x] = E[d_x] .$$

This immediately gives rise to a central limit theorem:

$$\frac{d_x - l_x v_x}{r_x} \,\tilde{\approx}\, N\left(0, \frac{E[d_x]}{r_x^2}\right),$$

where r_x is the total number of individuals observed at some time during age x. So

$$\frac{r_x}{v_x}\frac{d_x - l_x v_x}{r_x} \,\tilde{\approx}\, N\left(0, \frac{E[d_x]}{v_x^2}\right)$$

or

$$\hat{l}_x \mathrel{\dot\sim} N\left(l_x, \frac{E[d_x]}{v_x^2} \right).$$

Finally using the fact that

$$\begin{aligned} \frac{E[d_x]}{v_x} &= \frac{l_x E[v_x]}{v_x} \\ &\approx l_x \end{aligned}$$

it follows that

$$\hat{l}_x \mathrel{\dot\sim} N\left(l_x, \frac{l_x}{v_x} \right)$$

or a further approximation is

$$\hat{l}_x \mathrel{\dot\sim} N\left(l_x, \frac{d_x}{v_x^2} \right).$$

So in summary we assumed a piecewise constant hazard for each time interval. We derived the MLE and an approximation to the distribution. The resulting step function for the estimated hazard is $\hat{l}(t) = \hat{l}_x = d_x / v_x$ for $x \le t < x + 1$. The estimated survival distribution can be constructed using

$$\begin{aligned} \hat{S}(t) &= \exp\left(-\int_0^t \hat{l}(t) dt \right) \\ &= \prod_{i=0}^{x-1} e^{-\hat{l}_i} \end{aligned}$$

for $t = x$. Other methods of using the $\hat{l}_x$ are to assume that it is the hazard at the centre of the interval and derive a 'smooth' or 'graduation' of the points.

Example 5.1
Suppose we observed 10 individuals at some stage over the first 3 years of their life. Use the methods above to estimate the hazard in the second year of life and give an estimate of the standard error. If a and b are the ages of entry and Type I censoring in the first 3 years of life, then the data is

Individual	T	a	b
1	2.3	1.7	2.3
2	1.2	0	3
3	1.5	1.1	1.5
4	.5	0	3
5	1.6	0	3
6	2.1	0	3
7	.6	0	2
8	3	0	3
9	2.4	1.5	2.4
10	.6	0	1

To answer the problem, we need only consider individuals who were actually observed in their second year of life. Note that we will need to tabulate a_1 and b_1 for each individual. The relevant data are:

Individual	T_1	a_1	b_1	v_{1i}	d_{1i}
1	1	.7	1	.3	0
2	.2	0	1	.2	1
3	.5	.1	.5	.4	0
5	.6	0	1	.6	1
6	1	0	1	1	0
8	1	0	1	1	0
9	1	.5	1	.5	0

Summing the last two columns we obtain $d_1 = 2$, $v_1 = 4$ and so $\hat{l}_1 = .5$ and $SE\left[\hat{l}_1\right] = \sqrt{2/4^2} = \sqrt{2}/4$.

6. Multistate Markov Models

Now we wish to consider situations such as the one presented in figure 5.1. Notation is a serious issue for multistate models. We need a notation for the probability that an individual in state g at time x will move to state h at time $x+t$. There are various notations in common usage, for example $p_{g,h}(x, x+t)$ or ${}_t p_x^{gh}$. The former notation is more widely used by statisticians, the latter by actuaries. The actuarial notation will be used in this section. The same comments apply to the hazard function or rate of transitions from state g to state h at time t. The two most common notations for this quantity are $l_{g,h}(t)$ or l_t^{gh} with the latter being the notation used in this section.

A large amount of the material from the last section carries over immediately to this section. In particular all of the theory concerning inference for $l(t) = l_t$ is directly applicable. Define

$v^{gh}_{x,i}$ to be the amount of time that individual i during age x spends in state g at risk of making transitions to state h and $y^{gh}_{x,i}$ be the number of transitions actually made by the individual. Then the marginal likelihood of each transition by the individual is the same as (5.2). Under the piecewise exponential hazard assumption, the same analysis as in section 5 follows and we obtain the same expressions for the estimate of the transition intensities and standard errors:

$$\hat{l}^{gh}_{x} = \frac{y^{gh}_{x}}{v^{gh}_{x}} \text{ and } \hat{l}^{gh}_{x} \sim N\left(l^{gh}_{x}, \frac{y^{gh}_{x}}{\left(v^{gh}_{x}\right)^2} \right), \qquad (6.1)$$

where leaving out the i subscript in for example $y^{gh}_{x,i}$ implies summing over the individuals exposed to transitions from g to h in the age group. Also recall that the exact mean and variance may be obtained as

$$E\left[y^{gh}_{x} - l^{gh}_{x} v^{gh}_{x}\right] = 0\,,\ Var\left[y^{gh}_{x} - l^{gh}_{x} v^{gh}_{x}\right] = E\left[y^{gh}_{x}\right].$$

So the work done in the last section is immediately applicable and we are able to estimate l^{gh}_{x}. One further property (not proved here) is that

$$Cov\left[y^{gh}_{x} - l^{gh}_{x} v^{gh}_{x}, y^{rs}_{x} - l^{rs}_{x} v^{rs}_{x}\right] = 0$$

unless both $g = r$ and $h = s$. This means that it is fairly easy to derive multivariate central limit theorems which apply to a collection of estimates of distinct transition intensities. The mean vector will have entries given by (6.1), and the variance matrix will have diagonal entries also given by (6.1). Also note that, even though we have considered the age interval x to $x+1$, the results apply to age intervals of arbitrary length or whole of life, so long as it can be assumed that the transition intensities are constant.

Now that the estimation of l^{gh}_{x} has been resolved, we need to consider how we can use the estimates to construct estimates of other quantities of interest such as transition probabilities. We need a little more notation. Recall that we had previously defined ${}_{t}p^{gh}_{x}$ to be the probability that an individual in state g at time x will move to state h at time $x+t$. This does not say anything about the path taken and does not for example rule out multiple visits to g or h in the interval. A related quantity is the probability that the individual remains in g over the whole interval which is denoted by ${}_{t}p_{x}^{\overline{gg}}$. The probabilities ${}_{t}p_{x}^{gh}$ and ${}_{t}p_{x}^{\overline{gg}}$ can be shown to satisfy the Kolmogorov forward equations

$$\frac{\partial}{\partial t}{}_{t}p_{x}^{gh} = \sum_{r\neq h}\left({}_{t}p_{x}^{gr}\, l_{x+t}^{rh} - {}_{t}p_{x}^{gh}\, l_{x+t}^{hr}\right). \qquad (6.2)$$

and

$$\frac{\partial}{\partial t}{}_{t}p_{x}^{\overline{gg}} = -{}_{t}p_{x}^{\overline{gg}}\sum_{r\neq g} l_{x+t}^{gr}\ . \qquad (6.3)$$

In the usual way, we can show from (6.3) that

$${}_{t}p_{x}^{\overline{gg}} = \exp\left(-\int_{0}^{t}\sum_{r\neq g} l_{x+s}^{gr}\,ds\right). \qquad (6.4)$$

Obviously estimates of the transition intensities can be plugged into (6.4) to give estimates of ${}_{t}p_{x}^{\overline{gg}}$. Similarly estimates of the transition intensities can be plugged into (6.2). Often however (6.2) has no explicit solution and must be solved numerically, which is not difficult.

One practical issue when using (6.1) is that we need to know v_{x}^{gh} which is the time the individual is in state g able to make transitions to state h during age x. This quantity may not be available since the individual may not be observed continuously in many practical applications. In such cases it is usual to make assumptions such as deaths occur at the midpoint of the time interval on average.

Example 6.1
(Hypothetical). As a project we looked at the first year cohort of students. Eight hundred entered the faculty in 1999. Many students were in counselling during the year. The number of commencing sessions of counselling (so if students start, stop and restart counselling then they would count more than once) was 240 and the number of counselling sessions which ended was 140. The total time in counselling for the whole cohort in the year was 20 years and the total time not in counselling was 720 years. During the year 60 students in counselling and 40 students not in counselling terminated their course. Estimate the transition intensities between the states

1. N (not in counselling),
2. C (in counselling) and
3. T (terminated course).

Estimate the probability and a 95% confidence interval that an entering student will not seek counselling or terminate in the year.

Solution:
The stated information tells us that $v^{NC} = v^{NT} = 720$, $v^{CN} = v^{CT} = 20$, $y^{NC} = 240$, $y^{CN} = 140$, $y^{NT} = 40$ and $y^{CT} = 60$. So the estimates of the transition intensities are

$$\hat{l}^{NC} = y^{NC}/v^{NC} = 240/720 = 1/3\ ,$$

$$\hat{l}^{CN} = y^{CN}/v^{CN} = 140/20 = 7\ ,$$

$$\hat{l}^{NT} = y^{NT}/v^{NT} = 40/720 = 1/18\ ,$$

$$\hat{l}^{CT} = y^{CT}/v^{CT} = 60/20 = 3\ .$$

We are also asked to estimate $p^{\overline{NN}}$. From (6.4),

$$\begin{aligned}\hat{p}^{\overline{NN}} &= \exp\left(-\hat{l}^{NC} - \hat{l}^{NT}\right)\\ &= \exp\left(-1/3 - 1/18\right)\\ &= e^{-7/18}\\ &= .678.\end{aligned}$$

Since $\hat{l}^{NC}$ and $\hat{l}^{NT}$ are independent the approximate variance of $\hat{l}^{NC} + \hat{l}^{NT}$ is $y^{NC}/\left(v^{NC}\right)^2 + y^{NT}/\left(v^{NT}\right)^2$ and hence by the d − method, an approximate standard error for $\hat{p}^{\overline{NN}}$ is

$$\hat{p}^{\overline{NN}}\sqrt{y^{NC}/\left(v^{NC}\right)^2 + y^{NT}/\left(v^{NT}\right)^2}\ ,$$

which in this case is $.678\sqrt{240/720^2 + 40/720^2} = .0158$. So a 95% confidence interval for $p^{\overline{NN}}$ is $.678 \pm .031$.

7. Binomial method for estimation of survival

The emphasis in the last section was on estimating $l(t)$, for $x \le t < x+1$. Survival estimates and other associated quantities flowed from the estimate of $l(t)$. In this section a different approach is used. The focus is on the estimation of $q_x = {}_1q_x$ which is the probability of death in the year of age x given survival to that age. The data is recorded in a different manner to the previous section.

Consider one individual. The information that is recorded is the interval of observation in the age group and whether the individual died in the interval. Recall that $x+a$ is the age when the individual comes under observation and $x+b$ is the age at which the individual would stop being observed or would be Type I censored (in that age group). Recall that d is an indicator variable, which is 1 if the individual is observed to die in $(x+a, x+b]$ and 0 otherwise. Note that we are not using the actual time of death, which was the variable V from the last section. Putting aside for the moment the issue of random censoring, in order to use likelihood theory, we need the likelihood or pdf of d . The relevant probabilities are:

- The probability that the individual survives, which is $P(d = 0) = 1 - {}_{b-a}q_{x+a}$.
- The probability that the individual dies, which is $P(d = 1) = {}_{b-a}q_{x+a}$.

So the individual's contribution to the likelihood is ${}_{b-a}q_{x+a}^{d}\left(1 - {}_{b-a}q_{x+a}\right)^{1-d}$ and the overall likelihood is

$$L = \prod_{i \in \Re_x} {}_{b_i - a_i}q_{x+a_i}^{d_i}\left(1 - {}_{b_i - a_i}q_{x+a_i}\right)^{1-d_i} , \qquad (7.1)$$

where $\Re_x$ is the set of individuals observed at some time during age x. We are now in a very similar situation to the last section. In order to proceed we will need to make assumptions concerning the distribution of the survival random variable over the age interval. Taking account of the likelihood (7.1), the easiest way to proceed is to assume a specific form for ${}_tq_x$. The following three expressions are used commonly in practice:

- Uniform: ${}_tq_x = t \times q_x$,
- Constant hazard: ${}_tq_x = 1 - e^{-l_x t}$,
- Balducci: ${}_{1-t}q_{x+t} = (1-t) \times q_x$.

These can be substituted into (7.1), which can then be maximized to obtain $\hat{q}_x$. It will normally be necessary to maximize (7.1) by numerical means. In some special cases the estimators can be found explicitly. For example suppose all individuals started at the beginning of the age group and that all censoring took place at the midpoint of the interval. Let w_x be the number of individuals censored in the interval. Then $a_i = 0$ for all individuals, $b_i = 1$ for uncensored individuals and

$b_i = 1/2$ for censored individuals. So under the constant hazard assumption, for censored individuals we have

$$\begin{aligned} 1 - {}_{1/2}q_x &= 1 - \left(1 - e^{-\lambda_x/2}\right) \\ &= e^{-\lambda_x/2} \\ &= \left(1 - q_x\right)^{1/2}. \end{aligned}$$

So (7.1) becomes

$$q_x^{d_x}\left(1-q_x\right)^{r_x-d_x-w_x}\left(1-q_x\right)^{w_x/2} = q_x^{d_x}\left(1-q_x\right)^{\left(r_x-w_x/2\right)-d_x}, \quad (7.2)$$

where w_x is the number of individuals censored in the interval. The expression (7.2) is maximized by the 'actuarial estimator'

$$\hat{q}_x = \frac{d_x}{r_x - w_x/2}.$$

We will explore these models further in the tutorials.

8. Poisson method for estimation of survival

Yet another method for estimating survival uses the Poisson model for the number of events which occur during some interval of time. The method adds up the time over all individuals that an event (usually death) could have been observed. This is called the central exposed to risk and is denoted E_x^c. In the notation of section 5, this is the same as v_x. The Poisson model for the number of events d in an interval of length t is

$$P(d = j) = \frac{\lambda^j e^{-\lambda}}{j!}.$$

The key properties of a Poisson model are that the mean and variance of d are both λ and the maximum likelihood estimator of λ is d. For moderate and large λ, the Poisson distribution is well approximated by a normal with the same mean and variance. The Poisson method for estimating survival assumes that the number of events observed is Poisson with mean $\lambda_x E_x^c$, so

$$P(d_x = j) = \frac{\left(\lambda_x E_x^c\right)^j e^{-\lambda_x E_x^c}}{j!}.$$

So the Poisson maximum likelihood estimate of l_x is $\hat{l}_x = d_x \big/ E_x^c$. Note that this estimate is identical to the estimate in section 5. The motivation is different however.

Example 8.1

A population of university students was followed for 3 years to study the dropout rate. The population remained almost constant at about 500. There were 15 dropouts in that time. Estimate the dropout rate and give a 95% confidence interval.

Solution:
The central exposed to risk is 1500. So the estimated dropout rate is 1/100. Using the normal approximation to the Poisson, the approximate distribution of $\hat{l}$ is $N\left(l, l \big/ E^c\right)$. So a 95% confidence interval is

$$\begin{aligned}\hat{l} \pm 1.96\sqrt{\hat{l} \big/ E^c} &= .01 \pm 1.96\sqrt{.01/1500} \\ &= .01 \pm .00506.\end{aligned}$$

AUSTRALIAN NATIONAL UNIVERSITY

SCHOOL OF FINANCE AND APPLIED STATISTICS

SURVIVAL MODELS 1 (STAT3032/STAT8042) BIOSTATISTICS (STAT8003) SURVIVAL MODELS 1 (STAT3032/STAT8042) BIOSTATISTICS (STAT8003)

Part 3

Sections 9 and 10

9. Hypothesis Testing

The course so far has concentrated on estimation. We now wish to consider in more detail the testing of hypothesis concerning the survival experience that we have estimated. First we will recall some basic theory concerning statistical tests:

- Null hypothesis H_0: The null hypothesis usually represents the currently accepted view of the world. We need to have compelling evidence to renounce the status quo.
- Alternative hypothesis H_1: The alternative hypothesis is a different model for the real world which contradicts the null hypothesis. The evidence needs to be strong for the alternative to be accepted.
- The test statistic T is the summary of the relevant data which is used to quantify relevant information concerning the hypotheses and choose between the null and alternative hypotheses.
- The critical region C is the set of values of T which would lead us to reject the null hypothesis.
- The size of the test is the probability that T is in C when the null hypothesis is true.
- The power of the test is the probability that T is in C when the null hypothesis is false.

The properties of the test are determined by $P(T \in C)$ and hence the distribution of T. For most of the test statistics used in survival modeling, the exact distribution of the test statistics cannot be found. Instead, we must rely on asymptotic distributions, which apply when the sample sizes are sufficiently large. The two main distributions which arise in 'large sample' hypothesis testing are the normal and c^2 distributions which give rise to the 'z test' and the 'c^2 test'.

The general form of the 'z test' is a test statistic of the form $z = (T - m_0)/SE(T)$ where m_0 is the mean of T under the null hypothesis and $SE(T)$ is an estimate of the standard error of T. The resulting test can be one or two-tailed and may involve a 'continuity correction'.

The general form of the 'c^2 test' is a test statistic of the form

$$c^2 = \sum \frac{(O - E)^2}{E}$$

where 'O' stands for observed and 'E' for expected. One of the recurring issues with the 'c^2 test' is the degrees of freedom.

That depends on the number of terms in the summation and the number of parameters estimated to arrive at the E's. Any constraint on the observed such as for example a total being fixed reduces the degrees of freedom by one as does each parameter which has been estimated to arrive at the E's.

Example 9.1
Students who dropped out in 1999 were asked to give reasons. Historically we have a pretty good idea of why students leave. The data for 1999 was

Cause	Historical Percentage	Number in 1999	Expected number in 1999
No money	20	75	40
Too hard	50	65	100
Leaving town	15	30	30
Other causes	15	30	30

Test if the 1999 experience is consistent with the historical experience.

The c^2 test statistic is

$$\frac{(75-40)^2}{40}+\frac{(65-100)^2}{100}+\frac{(30-30)^2}{30}+\frac{(30-30)^2}{30}=42.75$$

Since the total of the E's is the same as the total of the O's, the degrees of freedom are reduced by one. So the test statistic is an observation from a $c^2(3)$ under the null hypothesis. It is highly significant and we conclude that the 99 result is different from the historical.

Some issues with the c^2 test are:

- All members of the group contributing to a specific 'O' term should be sufficiently similar so that the 'O' is approximately normal.
- If the members can be assumed to be homogeneous, then a rough rule of thumb is that each group should contain at least five members. Then the distribution of the overall statistic is likely to be well approximated by a c^2 distribution.
- The c^2 test us an omnibus test - as long as there is a difference between at least one of the E's and the O's the test will ultimately detect the difference.
- The flip side of the previous comment is that the c^2 test can have very low power against specific alternatives - designing a test that will work for any alternative can lead to a test which compares poorly to a Z test for a

specific alternative. For example suppose the groups correspond to age groups and we suspect that the study population has consistently higher mortality than the population standard. Then a much more powerful than the c^2 can be based on $\sum (O-E)$ which will have an approximate normal distribution. In general, by judicious choice of weights f in the weighted sum $\sum f(O-E)$, tests with high power can be designed for specific alternatives.

Example 9.2: Poisson
When we considered the Poisson model in the previous section we obtained the estimator $\hat{l}_x = d_x / E_x^C$, where we have previously defined the total time at risk in the age interval $[x, x+1)$ to be E_x^C . Since the hazard has been assumed to be constant across the interval, it is often more common to regard $\hat{l}_x$ as an estimate of the hazard at the centre of the interval, $l_{x+1/2}$ and write it as $\hat{l}_{x+1/2}$ instead. We now wish to compare the observed rates with a set of hypothesised or standard rates $l_{x+1/2}^S$. If the hypothesised rates are correct, then d_x will be approximately $N\left(E_x^C l_{x+1/2}^S, E_x^C l_{x+1/2}^S\right)$. So a Z test can be based on the test statistic,

$$Z = \frac{\sum_x f_x \left(d_x - E_x^C l_{x+1/2}^S\right)}{\sqrt{\sum_x f_x^2 E_x^C l_{x+1/2}^S}}, \qquad (9.1)$$

where the f_x are a suitably chosen set of weights. The statistics

$$Z_x = \frac{d_x - E_x^C l_{x+1/2}^S}{\sqrt{E_x^C l_{x+1/2}^S}}$$

may be used in diagnostic tests.

Example 9.4: Binomial
We can repeat the previous example using Binomial data. Recall that we had previously derived the actuarial estimator of q_x using the assumption that censoring if any occurs at the midpoint of the interval:

$$\hat{q}_x = \frac{d_x}{r_x - w_x/2}.$$

If the exact times of censoring b_i are known, then the denominator above can be replaced by the initial exposed to risk:

$$E_x = \sum (1-a_i) - \sum_{\text{non deaths}} (1-b_i),$$

where the first summation is over all individuals observed in age interval $(x, x+1)$ and the second summation is over the same age group but is restricted to those individuals who do not die. Then we can heuristically state that d_x is approximately Binomial(E_x, q_x^s) where q_x^s is a hypothesised or standard set of hazard rates for the age intervals $(x, x+1)$. Alternatively, we could write that d_x is approximately Normal$\left(E_x q_x^s, E_x q_x^s (1-q_x^s)\right)$. So a Z test can be based on the test statistic,

$$Z = \frac{\sum_x f_x \left(d_x - E_x q_x^s\right)}{\sqrt{\sum_x f_x^2 E_x q_x^s (1-q_x^s)}}, \qquad (9.2)$$

where the f_x are a suitably chosen set of weights. The statistics

$$Z_x = \frac{d_x - E_x q_x^s}{\sqrt{E_x q_x^s (1-q_x^s)}}$$

can be used in many diagnostic tests. A selection of tests are:

- c^2 **test**: $\sum_x Z_x^2$ is a c^2 test statistic. If m is the number of ages in the summation, then the degrees of freedom of the c^2 test is m minus the number of parameters that have been estimated. You will be asked in a tutorial to show that for the Binomial, $\sum_x Z_x^2$ can be rewritten in the form $\sum (O-E)^2 / E$.
- **Standardised deviations test**: Under the null hypothesis, the Z_x are a set of m (approximately) standard normal random variables and so tests can be based on comparing the set of Z_x to a standard normal. This can be done in a formal manner by binning the Z_x and comparing observed numbers to expected numbers using a c^2 test. For example the expected numbers of Z_x falling into the intervals $(-\infty, -1.96]$, $(-1.96, 0]$,

$(0,1.96]$ and $(1.96,\infty)$ are $.025m$, $.475m$, $.475m$ and $.025m$ respectively. Alternatively, the comparison to the normal could be done using a QQ plot.

- **Signs test**: The Z_x can be used in sign test. Under the null hypothesis the Z_x are equally likely to be positive or negative. So the number of positive Z_x is Binomial with parameters $(m,1/2)$ under the null hypothesis. The test is completed by finding the p-value of the observed number of positive values and checking against the specified significance level, for example .05.
- **Cumulative deviations:** The cumulative deviations test statistic is a special case of the general test statistics which were proposed in (9.1) and (9.2), with the weights being set to $f_i = 1$. This test would have good power for the alternative that mortality is either consistently too high or too low. So, for example, the test statistic for Poisson data would be

$$Z = \frac{\sum_x \left(d_x - E_x^C l_{x+1/2}^S\right)}{\sqrt{\sum_x E_x^C l_{x+1/2}^S}}.$$

- **Runs test:** This test looks at whether mortality tends to be higher than expected in some intervals and lower in others, rather than having the excesses and shortfalls randomly scattered over the age groups. Suppose that there are n_1 positive Z_x's and n_0 negative Z_x's. We would hope that the positive values are randomly scattered through the m age groups. One way of assessing the scatter of the positive values is to count the number of groups of positive values, t say. For example if the Z_x's were
{1.1, -1.2, -1.4, 2.4, 3, .5, 1.2, 2.1, -2.3, -1.1}
then we would have $m = 10$, $n_1 = 6$, $n_0 = 4$ and $t = 2$ positive groups. With this dataset, the maximum possible value of t is 5 and the minimum is 1. The issue is whether the observed value of 2 is surprising. Note that we should be surprised by small rather than large values of t. Also note that the minimum possible value of t will always be one corresponding to the arrangement which has either all positives first or last. We can use combinatorics to decide how close t must be to 1 to be significant. The number of ways of placing n_1 identical objects into m slots is $\binom{m}{n_1}$. We now need

to determine how many of those choices give t positive groups. We do it in two stages. First we note that the number of ways of dividing n_1 identical objects into t groups is $\binom{n_1-1}{t-1}$. One way to see this is to see this is to imagine the n_1 objects in a row and the groups are formed by inserting $t-1$ dividers between the objects. So there are n_1-1 possibilities for the locations of dividers and we need to choose $t-1$ of them. Now we need the numbers of ways of placing the t positive groups among the negative Z_x's. Think of the n_0 negative Z_x's in a line and we now need to choose the locations for the positive groups. A positive group can go between two negative values or at either end. So there are n_0+1 possible locations and t must be chosen giving $\binom{n_0+1}{t}$ possible choices. So the probability of getting exactly t groups is

$$\binom{n_1-1}{t-1}\binom{n_0+1}{t}\bigg/\binom{m}{n_1},$$

and the probability of getting t or less positive groups is

$$\sum_{j=1}^{t}\binom{n_1-1}{j-1}\binom{n_0+1}{j}\bigg/\binom{m}{n_1}.$$

In our example this is

$$\left(\binom{5}{0}\binom{5}{1}+\binom{5}{1}\binom{5}{2}\right)\bigg/\binom{10}{6}=\frac{11}{42},$$

which is not particularly surprising. So in this example, we would not conclude that there is clumping. For large m the distribution of t is approximately normal with mean $n_1(n_0+1)/(n_0+n_1)$ and variance $(n_0n_1)^2/(n_0+n_1)^3$.

- **Serial correlations:** Another way to test for clumping is to look for serial correlations or autocorrelations. The idea is that if there is no clumping or dependence between adjacent ages then Z_x and Z_{x+1} should be independent. So the true correlation coefficient between Z_x and Z_{x+1} should be 0. Suppose the ages under study are $1,\ldots,m$. Then we could estimate the correlation coefficient for (Z_x, Z_{x+1}), $x = 1, \ldots, m-1$ in the usual way, that is in the same way we estimate the correlation coefficient for n pairs of observations (X_i, Y_i). Indeed we could do the same for the j^{th} lagged sequences (Z_x, Z_{x+j}), $x = 1, \ldots, m-j$ to look for a correlation between observations that are j apart. The resulting correlation coefficient is called the j^{th} serial correlation coefficient. A commonly used approximation to the j^{th} serial correlation coefficient is given by $r_j \approx \left(\sum_{x=1}^{m-j}(z_x - \bar{z})(z_{x+j} - \bar{z})\right) \Big/ \left(\frac{m-j}{m}\sum_{x=1}^{m}(z_x - \bar{z})^2\right)$, but it is usually just as easy to find the true value. To complete the test, we need to know the distribution of the j^{th} serial correlation coefficient under the null hypothesis of zero correlation. An asymptotic approximation is usually used which says that the distribution of r_j is approximately normal with zero mean and variance $1/(m-j)$. So $r_j\sqrt{m-j}$ (or sometimes $r_j\sqrt{m}$) can be checked against the standard normal tables.

The above tests will be applied to a variety of examples in a forthcoming tutorial.

10. Smoothing

Scatterplot smoothing is a very common problem. We can treat it as a statistical problem or simply a problem in numerical analysis. For example, SPlus has a dataframe called air which can be used to explore the relationship between ozone concentration in the air and temperature. The help file description is:

SUMMARY

A data frame with 111 observations (rows), and 4 variables (columns), taken from an environmental study that measured the four variables ozone, solar radiation, temperature, and wind speed for 111 consecutive days.

"DATA DESCRIPTION

ozone	surface concentration of ozone in New York, in parts per million.
radiation	solar radiation
temperature	observed temperature, in degrees Fahrenheit.
wind	wind speed, in miles per hour.

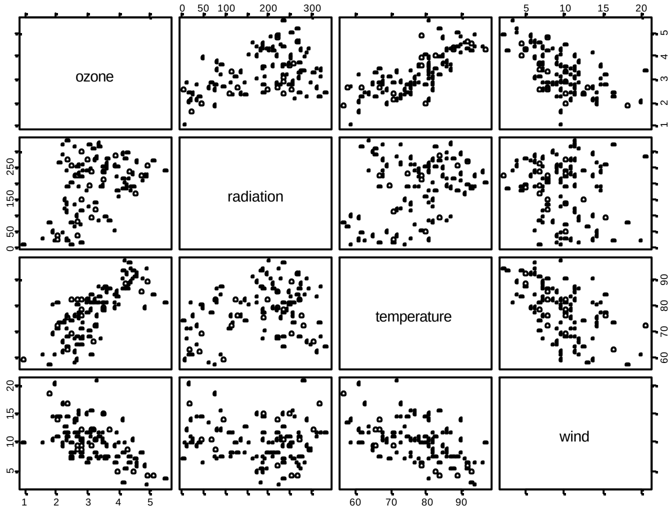
SOURCE

John M. Chambers and Trevor J. Hastie, (eds.) Statistical Models in S, Wadsworth and Brooks, Pacific Grove, CA 1992, pg. 348.

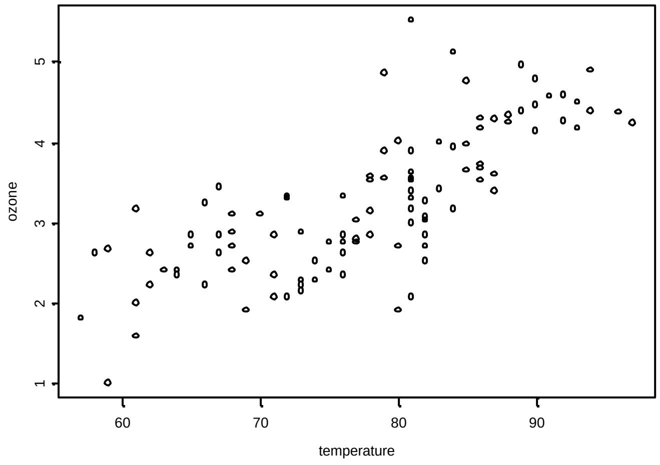
EXAMPLES

pairs(air)”

The pairs plot gives:



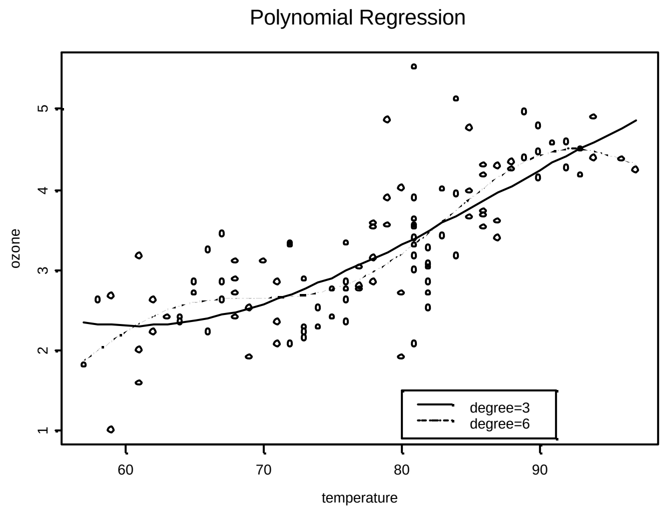
Let’s take a closer look at the relationship between ozone and temperature:



We are not sure of the relationship between ‘ozone’ and ‘temperature’ and would not necessarily wish to propose a linear relationship or indeed any specific functional form. A scatterplot smoother will draw a smooth curve against ‘ozone’. There are a variety of scatterplot smoothers available, many of which will give similar answers. Let’s try a few and then describe them later. The SPlus commands are:

```
> attach(air)
> #arrange the rows of air by increasing temperature
> newair_air[order(temperature),]
> detach("air")
> attach(newair)
> plot(temperature,ozone,main="Polynomial Regression")
> lines(temperature,fitted(lm(ozone~poly(temperature,3))))
>lines(temperature,fitted(lm(ozone~poly(temperature,6))),lty=3
+)
> legend(80,1.5,c("degree=3","degree=6"),lty=c(1,3))
```

The resulting plot is

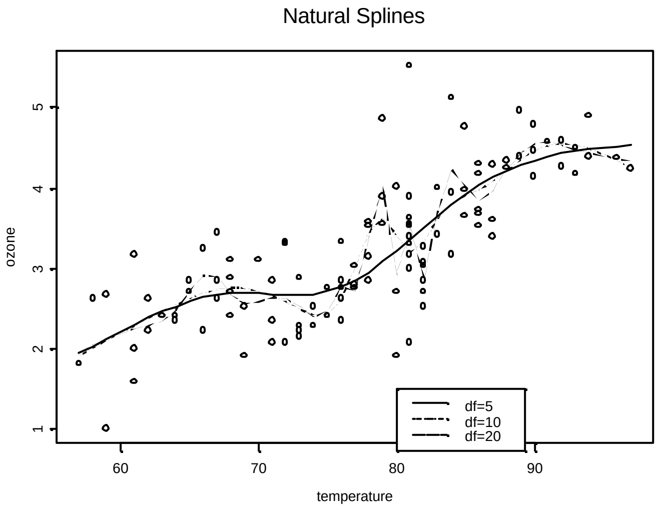


Now let’s try natural splines:

The SPlus instructions are:

```
> plot(temperature,ozone,main="Natural Splines")
> lines(temperature,fitted(lm(ozone~ns(temperature,5))))
> lines(temperature,fitted(lm(ozone~ns(temperature,10))),lty=3)
> lines(temperature,fitted(lm(ozone~ns(temperature,20))),lty=4)
> legend(80,1.5,c("df=5","df=10","df=20"),lty=c(1,3,4))
```

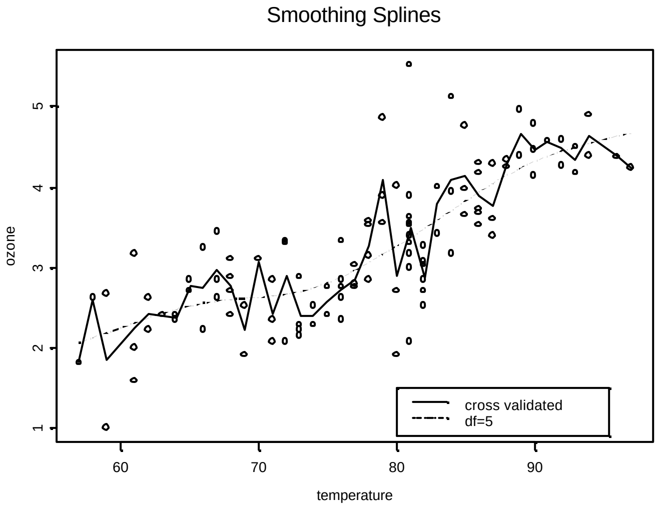
The resulting plot is:



Next try smoothing splines

```
> plot(temperature,ozone,main="Smoothing Splines")
> lines(smooth.spline(temperature,ozone))
> lines(smooth.spline(temperature,ozone,df=5),lty=3)
> legend(80,1.5,c("cross validated","df=5"),lty=c(1,3))
```

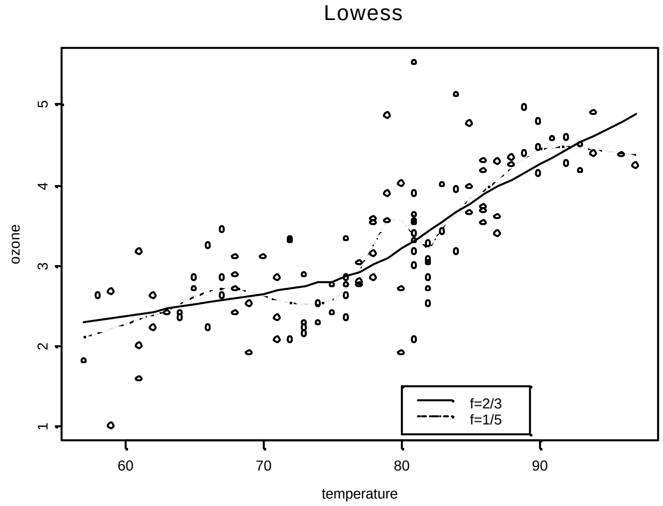
The resulting plot is



Note that the cross validated df (where the algorithm chooses df) was 38.52541. We will discuss this further later. Next try the Lowess smoother:

```
> plot(temperature,ozone,main="Lowess")
> lines(lowess(temperature,ozone))
> lines(lowess(temperature,ozone,f=.2),lty=3)
> #f is the fraction of the data used in the smooth at each point
> #the default is 2/3
> legend(80,1.5,c("f=2/3","f=1/5"),lty=c(1,3))
```

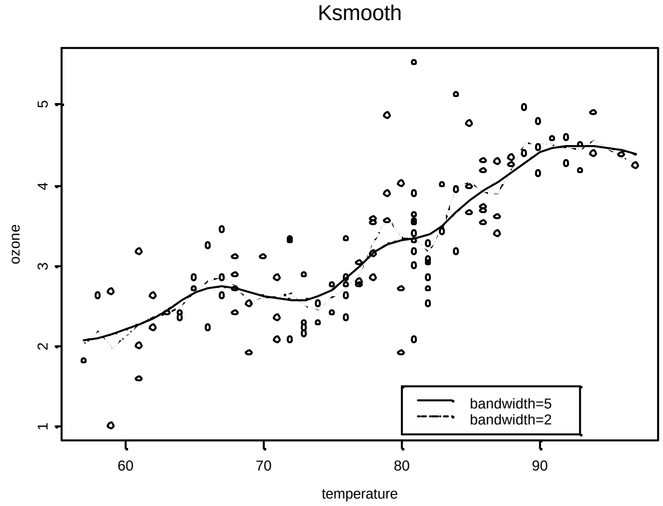
The resulting plot is :



Kernel smoothing is another popular type of smoothing:

```
> plot(temperature,ozone,main="Ksmooth")
>lines(ksmooth(temperature,ozone,kernel="normal",bandwidth
+=5))
>lines(ksmooth(temperature,ozone,"normal",bandwidth=2),lty=
+3)
> legend(80,1.5,c("bandwidth=5","bandwidth=2"),lty=c(1,3))
```

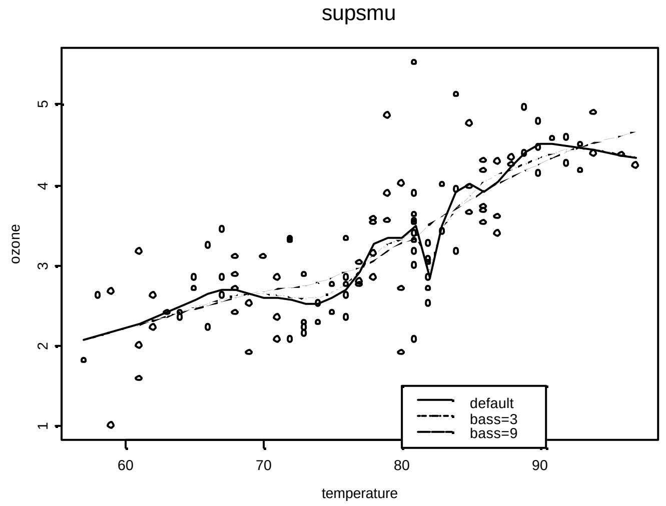
The resulting plot is:



A final smoother is 'Supersmoother':

```
> lines(supsmu(temperature,ozone))
> lines(supsmu(temperature,ozone,bass=3),lty=3)
> lines(supsmu(temperature,ozone,bass=9),lty=4)
> legend(80,1.5,c("default","bass=3","bass=9"),lty=c(1,3,4))
```

The resulting plot is:



We have seen that there are a variety of smoothers available, all of which can be used to smooth a scatterplot or do a 'graduation'. Given the sophistication of the smoothers, there is little place in modern data analysis for any hand based smoothing method. The main issue is how to choose between the multitude of smoothers. To assist in that choice, it is useful to try to get some idea of the key features of smoothers. Almost all smoothers will balance fidelity to the data with the roughness of the resulting curve. There is usually a parameter that adjusts the balance between fidelity and smoothness – it goes by various names some of which are bandwidth, degrees of freedom, span and smoothing parameter. The parameter can either be set by the user or obtained by minimizing an objective function over all possible choices of the parameter and the function. A common method of choosing the parameter is cross validation. The basic idea behind cross validation is to select a procedure using part of a sample and then see how it performs on the rest of the sample. The usual way that it is implemented in the smoothing area is as follows:

We wish to choose a scatterplot smoother for the dataset $\{(X_i, Y_i), i = 1, \ldots, n\}$. Suppose the parameter that governs the trade off between smoothness and fidelity is l and that the best smooth for a given value of l is $s(X|l)$. Then we would like to choose the value of l which minimizes

$$e^2(l) = E_{X,Y}(Y - s(X|l))^2 .$$

This quantity can only be estimated from the data. The method of estimation is cross validation. Let $s_{(i)}(x|l)$ be the smooth when the point (x_i, y_i) is left out of the data set and $s_{(i)}(x_i|l)$ be that smooth evaluated at the point x_i. We could then use $\left(y_i - s_{(i)}(x_i | l)\right)^2$ as an estimate of

$$e_{x_i}^2(l) = E_{Y|X=x_i}\left(Y - s(X|l)\right)^2 .$$

Finally, using the property that

$$\begin{aligned} e^2(l) &= E_{X,Y}\left(Y - s(X|l)\right)^2 \\ &= E_X\left\{E_{Y|X=x}\left(Y - s(x|l)\right)^2\right\} \\ &= E_X\left\{e_x^2(l)\right\} \end{aligned}$$

we can estimate $e^2(l)$ by

$$\hat{e}_{\mathrm{CV}}^2(l) = \sum_{i=1}^{n}\left(y_i - s_{(i)}(x_i|l)\right)^2 .$$

The cross validation estimator of l is the value of l which minimizes $\hat{e}_{\mathrm{CV}}^2(l)$.

We will now consider some of the more popular smoothers in a little more detail.

Kernel Smoothers:
In some ways the simplest smoothers are kernel smoothers. The smooth at a point for a kernel smoother is a simply weighted average of the local data points:

$$\hat{y}_i = \sum_{j=1}^{n} w_{ij} y_j \ ,$$

where the weights w_{ij}, which sum to 1, are given by

$$w_{ij} = \frac{K\left(\dfrac{x_i - x_j}{b}\right)}{\sum_{m=1}^{n} K\left(\dfrac{x_i - x_m}{b}\right)} .$$

The function K is called the kernel and the number b is called the bandwidth parameter (which can be chosen by cross validation as described above). The function K is often taken to be a continuous probability density function, very frequently the standard normal density. A good way to think about a kernel smooth at a data point is to imagine the underlying pdf centered at the data point and the weights for the other data points in the local average are proportional to the centered density. In general, the kernel K is usually assumed to have the following properties:

- $K(t) \geq 0$ for all t.
- $\int_{-\infty}^{\infty} K(t)dt = 1$.
- Symmetry: $K(t) = K(-t)$ for all t.

It is not necessary to use complicated kernels to get good results. For example the normal kernel

$$K(t) = \frac{e^{-t^2/2}}{\sqrt{2p}}$$

and the triangle kernel

$$K(t) = \begin{cases} 2 - 4|t|, & |t| \leq .5 \\ 0, & \text{otherwise} \end{cases}$$

give very similar results in practice. The latter may be preferred in practice because of its computational simplicity. Two other kernels that are implemented in SPlus are the "box" kernel

$$K(t) = \begin{cases} 1, & |t| \leq .5 \\ 0, & \text{otherwise} \end{cases}$$

and the Parzen kernel

$$K(t) = \begin{cases} \left(k_1 - t^2\right)/k_2, & |t| \leq C_1 \\ \left(t^2/k_3\right) - k_4|t| + k_5, & C_1 \leq |t| \leq C_2 \\ 0, & C_2 \leq |t| \end{cases}$$

The complexity of the Parzen kernel is not warranted in most practical applications.

When using a kernel smoother the impact of a point (x_j, y_j) on the smooth at x_i depends on how close x_j is to x_i, close values will lead to a large weight for y_j while far away points will get small or zero weight. The bandwidth parameter b determines how quickly $K(t/b)$ falls away and hence the width of $K(t/b)$. There is a trade off between bias and variance of the estimate, increasing the bandwidth lowers the variance of the smooth but increases the bias. One way of selecting the bandwidth is to minimize the mean square error of the smooth. The cross validation estimate of b minimizes an estimate of the mean square error. The choice of bandwidth has a much greater impact than the choice of kernel.

Natural splines or B-splines:
There are two functions in SPlus which can be used to fit splines. The functions approximate the scatterplot by piecewise polynomials. The default degree of the polynomials is 3, but there is a parameter that can be used to vary the degree – we won't do that. Knots are chosen to divide the horizontal axis into regions and distinct cubics are fitted in each region. The pieces are required to join smoothly, so at the knots, the cubics and their first and second derivatives agree. The smoothness is determined by the number of knots – more knots leads to a rougher curve. The knots can be supplied to the function or via df which lets the function choose the location of the df-1 internal knots. The two boundary knots are usually taken to be the outer limit of the data. Natural splines have the additional constraint that the fit is linear past the boundaries of the data or the boundary knots.

Smoothing Splines:
A smoothing spline behaves very similarly to a kernel smoother. Without worrying too much about the details, the smoothing spline approximation to the scatter plot is the function f with continuous and integrable first and second derivatives which minimizes

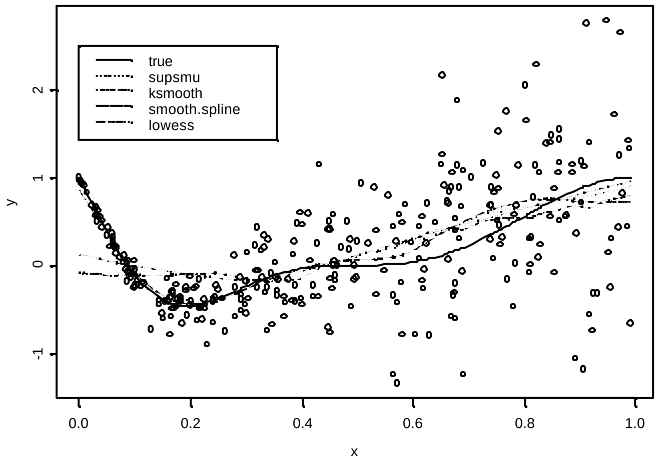
$$\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 + l\int\left(f''(x)\right)^2 dx$$

with l being the smoothing parameter. Larger values of l give smoother fits. If the smoothing parameter is not specified in SPlus then cross validation is used. Note that cross validation can give a smooth which is not particularly smooth. However it uses the estimate of l which minimizes the objective function.

Example 10.1
Consider constructing some data with a known signal and some noise and trying to recover the signal with the smoothers. Since the true curve is known we may get some insights about the behaviour of the smoothers. Suppose we add some noise to a sine curve:

```
> e_rnorm(300)
> x_runif(300)
> y_cos(2*pi*x)+sin(2*pi*(1-x)^2)+x*e
> plot(x,y)
> sort_order(x)
> x_x[sort]
> y_y[sort]
> fx_cos(2*pi*x)+sin(2*pi*(1-x)^2)
> lines(x,fx)
> lines(supsmu(x,y),lty=2)
> lines(ksmooth(x,y),lty=3)
> lines(smooth.spline(x,y),lty=4)
> lines(lowess(x,y),lty=5)
> legend(0,2.5,c("true","supsmu","ksmooth",
+ "smooth.spline","lowess"), lty=1:5)
```

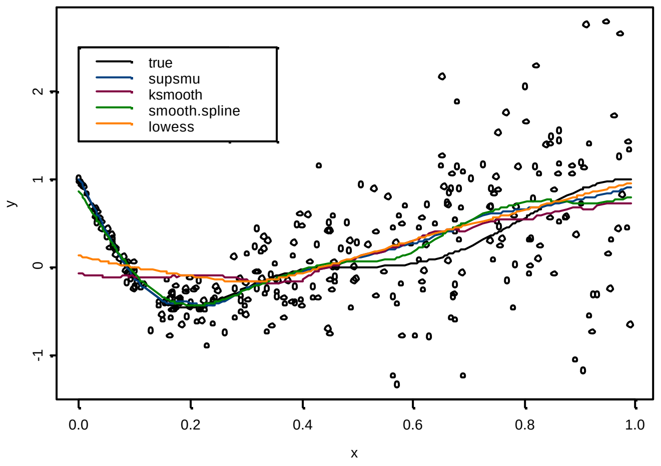
The resulting plot is



It is a little hard to distinguish the lines. Lets do the plot again with coloured lines. The lines will be easier to sort out on a colour screen or when printed on a colour printer, but will be hopeless when printed in black and white.

```
> e_rnorm(300)
> x_runif(300)
> y_cos(2*pi*x)+sin(2*pi*(1- x)^2)+x*e
> plot(x,y)
> sort_order(x)
> x_x[sort]
> y_y[sort]
> fx_cos(2*pi*x)+sin(2*pi*(1-x)^2)
> lines(x,fx,col=1)
> lines(supsmu(x,y),col=2)
> lines(ksmooth(x,y),col=3)
> lines(smooth.spline(x,y),col=4)
> lines(lowess(x,y),col=5)
> legend(0,2.5,c("true","supsmu","ksmooth",
+ "smooth.spline","lowess"), col=1:5,lty=1)
```

The resulting plot is



All of the smoothers do a very good job of recovering the signal in the centre of the data. Differences, if any, tend to occur at the boundaries of the data. A large part of the smoothing literature is concerned with boundary effects. On balance, the best smooth in the above example is probably supsmu. You should experiment with the smoothers in a variety of artificial examples like the above to get a feel for the behaviour of the smoothers.