Modern Applied Statistics with S

Fourth edition

by

W. N. Venables and B. D. Ripley

Springer (mid 2002)

Final 15 March 2002

Chapter 13

Survival Analysis

Extensive survival analysis facilities written by Terry Therneau (Mayo Foundation) are available in S-PLUS and in the R package `survival`.

Survival analysis is concerned with the distribution of lifetimes, often of humans but also of components and machines. There are two distinct levels of mathematical treatment in the literature. Collett (1994), Cox and Oakes (1984), Hosmer and Lemeshow (1999), Kalbfleisch and Prentice (1980) and Klein and Moeschberger (1997) take a traditional and mathematically non-rigorous approach. The modern mathematical approach based on continuous-parameter martingales is given by Fleming and Harrington (1991) and Andersen *et al.* (1993). (The latter is needed here only to justify some of the distribution theory and for the concept of martingale residuals.) Other aspects closely related to Therneau's software are described in Therneau and Grambsch (2000).

Let T denote a lifetime random variable. It will take values in $(0, \infty)$, and its continuous distribution may be specified by a cumulative distribution function F with a density f. (Mixed distributions can be considered, but many of the formulae used by the software need modification.) For lifetimes it is more usual to work with the *survivor function* $S(t) = 1 - F(t) = P(T > t)$, the *hazard function* $h(t) = \lim_{\Delta t \to 0} P(t \leqslant T < t + \Delta t \mid T \geqslant t)/\Delta t$ and the *cumulative hazard function* $H(t) = \int_0^t h(s)\,\mathrm{d}s$. These are all related; we have

$$h(t) = \frac{f(t)}{S(t)}, \qquad H(t) = -\log S(t)$$

Common parametric distributions for lifetimes are (Kalbfleisch and Prentice, 1980) the exponential, with $S(t) = \exp -\lambda t$ and hazard λ, the Weibull with

$$S(t) = \exp -(\lambda t)^\alpha, \qquad h(t) = \lambda\alpha(\lambda t)^{\alpha-1}$$

the log-normal, the gamma and the log-logistic which has

$$S(t) = \frac{1}{1 + (\lambda t)^\tau}, \qquad h(t) = \frac{\lambda\tau(\lambda t)^{\tau-1}}{1 + (\lambda t)^\tau}$$

The major distinguishing feature of survival analysis is *censoring*. An individual case may not be observed on the whole of its lifetime, so that, for example,

we may only know that it survived to the end of the trial. More general patterns of censoring are possible, but all lead to data for each case of the form either of a precise lifetime or the information that the lifetime fell in some interval (possibly extending to infinity).

Clearly we must place some restrictions on the censoring mechanism, for if cases were removed from the trial just before death we would be misled. Consider right censoring, in which the case leaves the trial at time C_i, and we know either T_i if $T_i \leqslant C_i$ or that $T_i > C_i$. *Random censoring* assumes that T_i and C_i are independent random variables, and therefore in a strong sense that censoring is uninformative. This includes the special case of *type I* censoring, in which the censoring time is fixed in advance, as well as trials in which the patients enter at random times but the trial is reviewed at a fixed time. It excludes *type II* censoring in which the trial is concluded after a fixed number of failures. Most analyses (including all those based solely on likelihoods) are valid under a weaker assumption that Kalbfleisch and Prentice (1980, §5.2) call *independent* censoring in which the hazard at time t conditional on the whole history of the process only depends on the survival of that individual to time t. (Independent censoring does cover type II censoring.) Conventionally the time recorded is $\min(T_i, C_i)$ together with the indicator variable for observed death $\delta_i = I(T_i \leqslant C_i)$. Then under independent right censoring the likelihood for parameters in the lifetime distribution is

$$L = \prod_{\delta_i=1} f(t_i) \prod_{\delta_i=0} S(t_i) = \prod_{\delta_i=1} h(t_i)S(t_i) \prod_{\delta_i=0} S(t_i) = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i) \tag{13.1}$$

Usually we are not primarily interested in the lifetime distribution *per se*, but how it varies between groups (usually called *strata* in the survival context) or on measurements on the cases, called *covariates*. In the more complicated problems the hazard will depend on covariates that vary with time, such as blood pressure or changes of treatments.

The function `Surv(times, status)` is used to describe the censored survival data to the S functions, and always appears on the left side of a model formula. In the simplest case of right censoring the variables are $\min(T_i, C_i)$ and δ_i (logical or 0/1 or 1/2). Further forms allow left and interval censoring. The results of printing the object returned by `Surv` are the vector of the information available, either the lifetime or an interval.

We consider three small running examples. Uncensored data on survival times for leukaemia (Feigl and Zelen, 1965; Cox and Oakes, 1984, p. 9) are in data frame `leuk`. This has two covariates, the white blood count `wbc`, and `ag` a test result that returns 'present' or 'absent'. Two-sample data (Gehan, 1965; Cox and Oakes, 1984, p. 7) on remission times for leukaemia are given in data frame `gehan`. This trial has 42 individuals in matched pairs, and no covariates (other than the treatment group).[1] Data frame `motors` contains the results of an accelerated life test experiment with 10 replicates at each of four temperatures reported

[1] Andersen *et al.* (1993, p. 22) indicate that this trial had a sequential stopping rule that invalidates most of the methods used here; it should be seen as illustrative only.

by Nelson and Hahn (1972) and Kalbfleisch and Prentice (1980, pp. 4–5). The times are given in hours, but all but one is a multiple of 12, and only 14 values occur

17 21 22 56 60 70 73.5 115.5 143.5 147.5833 157.5 202.5 216.5 227

in days, which suggests that observation was not continuous. Thus this is a good example to test the handling of ties.

13.1 Estimators of Survivor Curves

The estimate of the survivor curve for uncensored data is easy; just take one minus the empirical distribution function. For the leukaemia data we have

```
plot(survfit(Surv(time) ~ ag, data=leuk), lty = 2:3, col = 2:3)
legend(80, 0.8, c("ag absent", "ag present"), lty = 2:3, col = 2:3)
```

and confidence intervals are obtained easily from the binomial distribution of $\widehat{S}(t)$. For example, the estimated variance is

$$\widehat{S}(t)[1 - \widehat{S}(t)]/n = r(t)[n - r(t)]/n^3 \tag{13.2}$$

when $r(t)$ is the number of cases still alive (and hence 'at risk') at time t.

This computation introduces the function `survfit` and its associated `plot`, `print` and `summary` methods. It takes a model formula, and if there are factors on the right-hand side, splits the data on those factors, and plots a survivor curve for each factor combination, here just presence or absence of `ag`. (Although the factors can be specified additively, the computation effectively uses their interaction.)

For censored data we have to allow for the decline in the number of cases at risk over time. Let $r(t)$ be the number of cases at risk just before time t, that is, those that are in the trial and not yet dead. If we consider a set of intervals $I_i = [t_i, t_{i+1})$ covering $[0, \infty)$, we can estimate the probability p_i of surviving interval I_i as $[r(t_i) - d_i]/r(t_i)$ where d_i is the number of deaths in interval I_i. Then the probability of surviving until t_i is

$$P(T > t_i) = S(t_i) \approx \prod_0^{i-1} p_j \approx \prod_0^{i-1} \frac{r(t_i) - d_i}{r(t_i)}$$

Now let us refine the grid of intervals. Non-unity terms in the product will only appear for intervals in which deaths occur, so the limit becomes

$$\widehat{S}(t) = \prod \frac{r(t_i) - d_i}{r(t_i)}$$

the product being over times at which deaths occur before t (but they could occur simultaneously). This is the Kaplan–Meier estimator. Note that this becomes constant after the largest observed t_i, and for this reason the estimate is only plotted up to the largest t_i. However, the points at the right-hand end of the plot

will be very variable, and it may be better to stop plotting when there are still a few individuals at risk.

We can apply similar reasoning to the cumulative hazard

$$H(t_i) \approx \sum_{j \leqslant i} h(t_j)(t_{j+1} - t_j) \approx \sum_{j \leqslant i} \frac{d_j}{r(t_j)}$$

with limit

$$\widehat{H}(t) = \sum \frac{d_j}{r(t_j)} \tag{13.3}$$

again over times at which deaths occur before t. This is the Nelson estimator of the cumulative hazard, and leads to the Altshuler or Fleming–Harrington estimator of the survivor curve

$$\tilde{S}(t) = \exp -\widehat{H}(t) \tag{13.4}$$

The two estimators are related by the approximation $\exp -x \approx 1 - x$ for small x, so they will be nearly equal for large risk sets. The S functions follow Fleming and Harrington in breaking ties in (13.3), so if there were 3 deaths when the risk set contained 12 people, $3/12$ is replaced by $1/12 + 1/11 + 1/10$.

Similar arguments to those used to derive the two estimators lead to the standard error formula for the Kaplan–Meier estimator

$$\text{var}\left(\widehat{S}(t)\right) = \widehat{S}(t)^2 \sum \frac{d_j}{r(t_j)[r(t_j) - d_j]} \tag{13.5}$$

often called Greenwood's formula after its version for life tables, and

$$\text{var}\left(\widehat{H}(t)\right) = \sum \frac{d_j}{r(t_j)[r(t_j) - d_j]} \tag{13.6}$$

We leave it to the reader to check that Greenwood's formula reduces to (13.2) in the absence of ties and censoring. Note that if censoring can occur, both the Kaplan–Meier and Nelson estimators are biased; the bias results from the inability to give a sensible estimate when the risk set is empty.

Tsiatis (1981) suggested the denominator $r(t_j)^2$ rather than $r(t_j)[r(t_j) - d_j]$ on asymptotic grounds. Both Fleming and Harrington (1991) and Andersen *et al.* (1993) give a rigorous derivation of these formulae (and corrected versions for mixed distributions), as well as calculations of bias and limit theorems that justify asymptotic normality. Klein (1991) discussed the bias and small-sample behaviour of the variance estimators; his conclusions for $\widehat{H}(t)$ are that the bias is negligible and the Tsiatis form of the standard error is accurate (for practical use) provided the expected size of the risk set at t is at least five. For the Kaplan–Meier estimator Greenwood's formula is preferred, and is accurate enough (but biased downwards) again provided the expected size of the risk set is at least five.

We can use these formulae to indicate confidence intervals based on asymptotic normality, but we must decide on what scale to compute them. By default the

function `survfit` computes confidence intervals on the log survivor (or cumulative hazard) scale, but linear and complementary log-log scales are also available (via the `conf.type` argument). These choices give

$$\widehat{S}(t)\exp\left[\pm k_\alpha \text{ s.e.}(\widehat{H}(t))\right]$$
$$\widehat{S}(t)\left[1 \pm k_\alpha \text{ s.e.}(\widehat{H}(t))\right]$$
$$\exp\Big\{-\widehat{H}(t)\exp\Big[\pm k_\alpha \frac{\text{s.e.}(\widehat{H}(t))}{\widehat{H}(t)}\Big]\Big\}$$

the last having the advantage of taking values in $(0, 1)$. Bie, Borgan and Liestøl (1987) and Borgan and Liestøl (1990) considered these and an arc-sine transformation; their results indicate that the complementary log-log interval is quite satisfactory for sample sizes as small as 25.

We do not distinguish clearly between log-survivor curves and cumulative hazards, which differ only by sign, yet the natural estimator of the first is the Kaplan–Meier estimator on log scale, and for the second it is the Nelson estimator. This is particularly true for confidence intervals, which we would expect to transform just by a change of sign. Fortunately, practical differences only emerge for very small risk sets, and are then swamped by the very large variability of the estimators.

The function `survfit` also handles censored data, and uses the Kaplan–Meier estimator by default. We try it on the `gehan` data:

```
> attach(gehan)
> Surv(time, cens)
 [1]  1  10  22   7   3  32+ 12  23   8  22  17   6   2  16
[15] 11  34+  8  32+ 12  25+  2  11+  5  20+  4  19+ 15   6
[29]  8  17+ 23  35+  5   6  11  13   4   9+  1   6+  8  10+
> plot(log(time) ~ pair)
> gehan.surv <- survfit(Surv(time, cens) ~ treat, data = gehan,
     conf.type = "log-log")
> summary(gehan.surv)
   ....
> plot(gehan.surv, conf.int = T, lty = 3:2, log = T,
     xlab = "time of remission (weeks)", ylab = "survival")
> lines(gehan.surv, lty = 3:2, lwd = 2, cex = 2)
> legend(25, 0.1 , c("control", "6-MP"), lty = 2:3, lwd = 2)
```

which calculates and plots (as shown in Figure 13.1) the product-limit estimators for the two groups, giving standard errors calculated using Greenwood's formula. (Confidence intervals are plotted automatically if there is only one group.) Other options are available, including `error = "tsiatis"` and `type = "fleming-harrington"` (which can be abbreviated to the first character). Note that the `plot` method has a `log` argument that plots $\widehat{S}(t)$ on log scale, effectively showing the negative cumulative hazard.

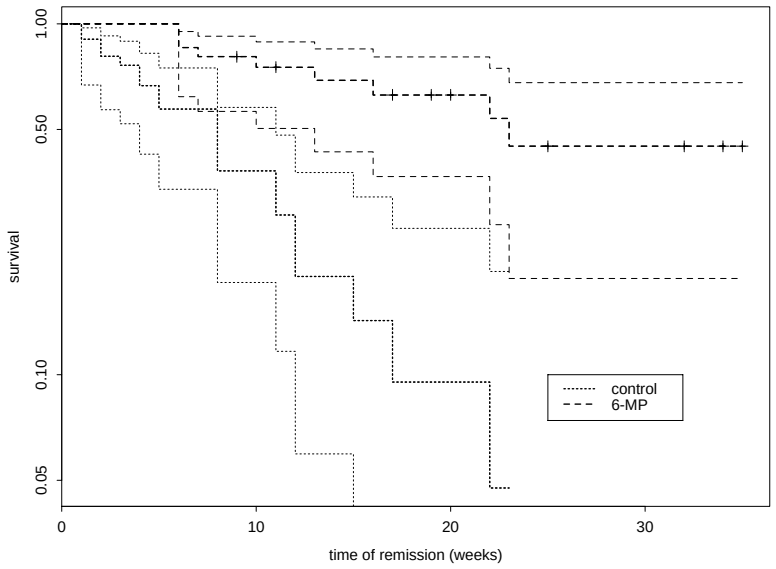

Figure 13.1: Survivor curves (on log scale) for the two groups of the `gehan` data. The crosses (on the 6-MP curve) represent censoring times. The thicker lines are the estimates, the thinner lines pointwise 95% confidence intervals.

Testing survivor curves

We can test for differences between the groups in the `gehan` example by

```
> survdiff(Surv(time, cens) ~ treat, data = gehan)
               N Observed Expected (O-E)^2/E (O-E)^2/V
   treat=6-MP 21        9     19.3      5.46      16.8
treat=control 21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```

This is one of a family of tests with parameter ρ defined by Fleming and Harrington (1981) and Harrington and Fleming (1982). The default $\rho = 0$ corresponds to the *log-rank test*. Suppose t_j are the observed death times. If we condition on the risk set and the number of deaths D_j at time t_j, the mean of the number of deaths D_{jk} in group k is clearly $E_{jk} = D_k r_k(t_j)/r(t_j)$ under the null hypothesis (where $r_k(t_j)$ is the number from group k at risk at time j). The statistic used is $(O_k - E_k) = \sum_j \widehat{S}(t_j-)^\rho \left[D_{jk} - E_{jk}\right]$,[2] and from this we compute a statistic $(O-E)^T V^{-1}(O-E)$ with an approximately chi-squared distribution. There are a number of different approximations to the variance matrix V, the one used being the weighted sum over death times of the variance matrices of $D_{jk} - E_{jk}$ computed from the hypergeometric distribution. The sum of $(O_k - E_k)^2/E_k$ provides a conservative approximation to the chi-squared statistic. The final column is $(O-E)^2$ divided by the diagonal of V; the final line gives the overall statistic computed from the full quadratic form.

[2] $\widehat{S}(t-)$ is the Kaplan–Meier estimate of survival just prior to t, ignoring the grouping.

The value `rho = 1` corresponds approximately to the Peto–Peto modification (Peto and Peto, 1972) of the Wilcoxon test, and is more sensitive to early differences in the survivor curves.

A warning: tests of differences between groups are often used inappropriately. The `gehan` dataset has no other covariates, but where there are covariates the differences between the groups may reflect or be masked by differences in the covariates. Thus for the `leuk` dataset

```
> survdiff(Surv(time) ~ ag, data = leuk)
            N Observed Expected (O-E)^2/E (O-E)^2/V
 ag=absent 16       16      9.3      4.83      8.45
ag=present 17       17     23.7      1.90      8.45

 Chisq= 8.4  on 1 degrees of freedom, p= 0.00365
```

is inappropriate as there are differences in distribution of `wbc` between the two groups. A model is needed to adjust for the covariates (see page 368).

13.2 Parametric Models

Parametric models for survival data have fallen out of fashion with the advent of less parametric approaches such as the Cox proportional hazard models considered in the next section, but they remain a very useful tool, particularly in exploratory work (as usually they can be fitted very much faster than the Cox models).

The simplest parametric model is the exponential distribution with hazard $\lambda_i > 0$. The natural way to relate this to a covariate vector $\boldsymbol{x}$ for the case (including a constant if required) and to satisfy the positivity constraint is to take

$$\log \lambda_i = \boldsymbol{\beta}^T \boldsymbol{x}_i, \qquad \lambda_i = e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}$$

For the Weibull distribution the hazard function is

$$h(t) = \lambda^\alpha \alpha t^{\alpha-1} = \alpha t^{\alpha-1} \exp(\alpha \boldsymbol{\beta}^T \boldsymbol{x}) \tag{13.7}$$

if we again make λ an exponential function of the covariates, and so we have the first appearance of the *proportional hazards* model

$$h(t) = h_0(t) \exp \boldsymbol{\beta}^T \boldsymbol{x} \tag{13.8}$$

which we consider again later. This identification suggests re-parametrizing the Weibull by replacing λ^α by λ, but as this just rescales the coefficients we can move easily from one parametrization to the other.

The Weibull is also a member of the class of *accelerated life* models, which have survival time T such that $T \exp \boldsymbol{\beta}^T \boldsymbol{x}$ has a fixed distribution; that is, time is speeded up by the factor $\exp \boldsymbol{\beta}^T \boldsymbol{x}$ for an individual with covariate $\boldsymbol{x}$. This corresponds to replacing t in the survivor function and hazard by $t \exp \boldsymbol{\beta}^T \boldsymbol{x}$, and for models such as the exponential, Weibull and log-logistic with parametric

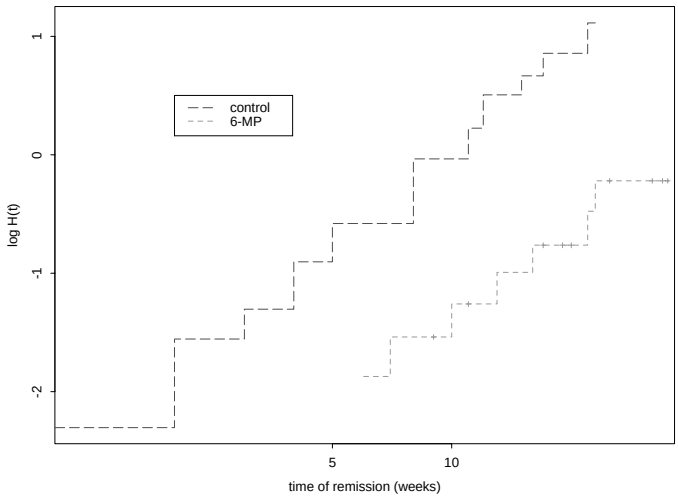

Figure 13.2: A log-log plot of cumulative hazard for the `gehan` dataset.

dependence on λt, this corresponds to taking $\lambda = \exp \boldsymbol{\beta}^T \boldsymbol{x}$. For all accelerated-life models we will have

$$\log T = \log T_0 - \boldsymbol{\beta}^T \boldsymbol{x} \tag{13.9}$$

for a random variable T_0 whose distribution does not depend on $\boldsymbol{x}$, so these are naturally considered as regression models.

For the Weibull the cumulative hazard is linear on a log-log plot, which provides a useful diagnostic aid. For example, for the `gehan` data

```
> plot(gehan.surv,  lty = 3:4, col = 2:3, fun = "cloglog",
       xlab = "time of remission (weeks)", ylab = "log H(t)")
> legend(2, 0.5, c("control","6-MP"), lty = 4:3, col = 3:2)
```

we see excellent agreement with the proportional hazards hypothesis and with a Weibull baseline (Figure 13.2).

The function `survReg`[3] fits parametric survival models of the form

$$\ell(T) \sim \boldsymbol{\beta}^T \boldsymbol{x} + \sigma\, \epsilon \tag{13.10}$$

where $\ell(\,)$ is usually a log transformation. The `dist` argument specifies the distribution of ϵ and $\ell(\,)$, and σ is known as the *scale*. The distribution can be `weibull` (the default) `exponential`, `rayleigh`, `lognormal` or `loglogistic`, all with a log transformation, or `extreme`, `logistic`, `gaussian` or `t` with an identity transformation.

The default for `distribution` corresponds to the model

$$\log T \sim \boldsymbol{\beta}^T \boldsymbol{x} + \sigma \log E$$

for a standard exponential E whereas our Weibull parametrization corresponds to

$$\log T \sim -\log \lambda + \frac{1}{\alpha} \log E$$

[3] `survreg` in R.

Thus `survReg` uses a log-linear Weibull model for $-\log\lambda$ and the scale factor σ estimates $1/\alpha$. The `exponential` distribution comes from fixing $\sigma = \alpha = 1$.

We consider exponential analyses, followed by Weibull and log-logistic regression analyses.

```
> options(contrasts = c("contr.treatment", "contr.poly"))
> survReg(Surv(time) ~ ag*log(wbc), data = leuk,
          dist = "exponential")
    ....
Coefficients:
 (Intercept)    ag log(wbc) ag:log(wbc)
      4.3433 4.135 -0.15402    -0.32781

Scale fixed at 1

Loglik(model)= -145.7   Loglik(intercept only)= -155.5
        Chisq= 19.58 on 3 degrees of freedom, p= 0.00021
> summary(survReg(Surv(time) ~ ag + log(wbc), data = leuk,
                  dist = "exponential"))
             Value Std. Error     z        p
(Intercept)  5.815      1.263  4.60 4.15e-06
         ag  1.018      0.364  2.80 5.14e-03
   log(wbc) -0.304      0.124 -2.45 1.44e-02

> summary(survReg(Surv(time) ~ ag + log(wbc), data = leuk))
# Weibull is the default
    ....
              Value Std. Error      z        p
(Intercept)  5.8524      1.323  4.425 9.66e-06
         ag  1.0206      0.378  2.699 6.95e-03
   log(wbc) -0.3103      0.131 -2.363 1.81e-02
 Log(scale)  0.0399      0.139  0.287 7.74e-01

Scale= 1.04

Weibull distribution
Loglik(model)= -146.5   Loglik(intercept only)= -153.6
        Chisq= 14.18 on 2 degrees of freedom, p= 0.00084
    ....
> summary(survReg(Surv(time) ~ ag + log(wbc), data = leuk,
                  dist = "loglogistic"))
             Value Std. Error     z        p
(Intercept)  8.027      1.701  4.72 2.37e-06
         ag  1.155      0.431  2.68 7.30e-03
   log(wbc) -0.609      0.176 -3.47 5.21e-04
 Log(scale) -0.374      0.145 -2.58 9.74e-03

Scale= 0.688

Log logistic distribution
```

```
Loglik(model)= -146.6   Loglik(intercept only)= -155.4
        Chisq= 17.58 on 2 degrees of freedom, p= 0.00015
```

The Weibull analysis shows no support for non-exponential shape. For later reference, in the proportional hazards parametrization (13.8) the estimate of the coefficients is $\widehat{\boldsymbol{\beta}} = -(5.85, 1.02, -0.310)^T/1.04 = (-5.63, -0.981, 0.298)^T$. The log-logistic distribution, which is an accelerated life model but not a proportional hazards model (in our parametrization), gives a considerably more significant coefficient for `log(wbc)`. Its usual scale parameter τ (as defined on page 353) is estimated as $1/0.688 \approx 1.45$.

We can test for a difference in groups within the Weibull model by the Wald test (the 'z' value for `ag`) or we can perform a likelihood ratio test by the `anova` method.

```
> anova(survReg(Surv(time) ~ log(wbc), data = leuk),
        survReg(Surv(time) ~ ag + log(wbc), data = leuk))
    ....
          Terms Resid. Df  -2*LL Test Df Deviance Pr(Chi)
1      log(wbc)        30 299.35
2 ag + log(wbc)        29 293.00  +ag  1   6.3572 0.01169
```

The likelihood ratio test statistic is somewhat less significant than the result given by `survdiff`.

An extension is to allow different scale parameters σ for each group, by adding a `strata` argument to the formula. For example,

```
> summary(survReg(Surv(time) ~ strata(ag) + log(wbc), data=leuk))
    ....
             Value Std. Error      z        p
(Intercept)  7.499      1.475  5.085 3.68e-07
   log(wbc) -0.422      0.149 -2.834 4.59e-03
  ag=absent  0.152      0.221  0.688 4.92e-01
 ag=present  0.142      0.216  0.658 5.11e-01

Scale:
 ag=absent ag=present
      1.16       1.15

Weibull distribution
Loglik(model)= -149.7   Loglik(intercept only)= -153.2
    ....
```

If the accelerated-life model holds, $T\exp(-\boldsymbol{\beta}^T\boldsymbol{x})$ has the same distribution for all subjects, being standard Weibull, log-logistic and so on. Thus we can get some insight into what the common distribution should be by studying the distribution of $(T_i \exp(-\widehat{\boldsymbol{\beta}}^T\boldsymbol{x}_i))$. Another way to look at this is that the residuals from the regression are $\log T_i - \widehat{\boldsymbol{\beta}}^T\boldsymbol{x}_i$ which we have transformed back to the scale of time. For the `leuk` data we could use, for example,

```
leuk.wei <- survReg(Surv(time) ~ ag + log(wbc), data = leuk)
ntimes <- leuk$time * exp(-leuk.wei$linear.predictors)
plot(survfit(Surv(ntimes)), log = T)
```

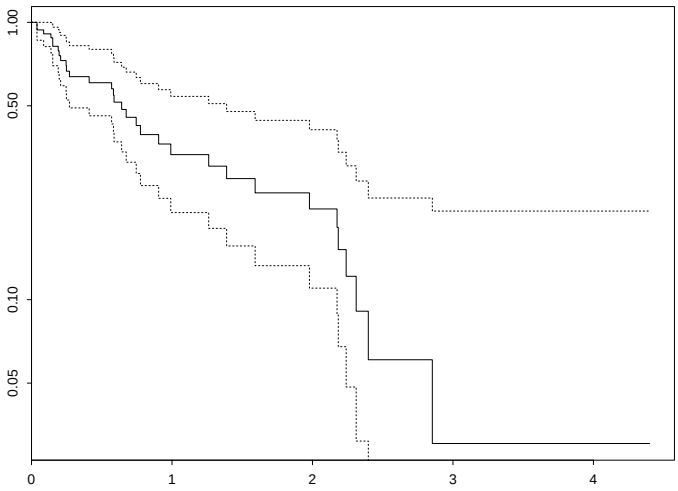

Figure 13.3: A log plot of $S(t)$ (equivalently, a linear plot of $-H(t)$) for the `leuk` dataset with pointwise confidence intervals.

The result (Figure 13.3) is plausibly linear, confirming the suitability of an exponential model. If we wished to test for a general Weibull distribution, we should plot $\log(-\log \widehat{S}(t))$ against $\log t$. (This is provided by the `fun="cloglog"` argument to `plot.survfit`)

Moving on to the `gehan` dataset, which includes right censoring, we find

```
> survReg(Surv(time, cens) ~ factor(pair) + treat, data = gehan,
          dist = "exponential")
    ....
Loglik(model)= -101.6   Loglik(intercept only)= -116.8
        Chisq= 30.27 on 21 degrees of freedom, p= 0.087
> summary(survReg(Surv(time, cens) ~ treat, data = gehan,
                  dist = "exponential"))
            Value Std. Error      z        p
(Intercept)  3.69      0.333  11.06 2.00e-28
      treat -1.53      0.398  -3.83 1.27e-04
Scale fixed at 1

Exponential distribution
Loglik(model)= -108.5   Loglik(intercept only)= -116.8
        Chisq= 16.49 on 1 degrees of freedom, p= 4.9e-05
> summary(survReg(Surv(time, cens) ~ treat, data = gehan))
             Value Std. Error      z        p
(Intercept)  3.516      0.252  13.96 2.61e-44
      treat -1.267      0.311  -4.08 4.51e-05
 Log(scale) -0.312      0.147  -2.12 3.43e-02

Scale= 0.732

Weibull distribution
Loglik(model)= -106.6   Loglik(intercept only)= -116.4
        Chisq= 19.65 on 1 degrees of freedom, p= 9.3e-06
```

There is no evidence of close matching of pairs. The difference in log hazard between treatments is $-(-1.267)/0.732 = 1.73$ with a standard error of $0.42 = 0.311/0.732$.

Finally, we consider the `motors` data, which are analysed by Kalbfleisch and Prentice (1980, §3.8.1). According to Nelson and Hahn (1972), the data were collected to assess survival at $130°\mathrm{C}$, for which they found a median of 34 400 hours and a 10 percentile of 17 300 hours.

```
> plot(survfit(Surv(time, cens) ~ factor(temp), data = motors),
       conf.int = F)
> motor.wei <- survReg(Surv(time, cens) ~ temp, data = motors)
> summary(motor.wei)
               Value Std. Error     z         p
(Intercept)  16.3185    0.62296  26.2 3.03e-151
       temp  -0.0453    0.00319 -14.2  6.74e-46
 Log(scale)  -1.0956    0.21480  -5.1  3.38e-07

Scale= 0.334

Weibull distribution
Loglik(model)= -147.4   Loglik(intercept only)= -169.5
        Chisq= 44.32 on 1 degrees of freedom, p= 2.8e-11
    ....
> unlist(predict(motor.wei, data.frame(temp=130), se.fit = T))
    fit se.fit
  33813 7506.3
```

The `predict` method by default predicts the centre of the distribution. We can obtain predictions for quantiles by

```
> predict(motor.wei, data.frame(temp=130), type = "quantile",
          p = c(0.5, 0.1))
[1] 29914 15935
```

We can also use `predict` to find standard errors, but we prefer to compute confidence intervals on log-time scale by

```
> t1 <-  predict(motor.wei, data.frame(temp=130),
                 type = "uquantile", p = 0.5, se = T)
> exp(c(LL=t1$fit - 2*t1$se, UL=t1$fit + 2*t1$se))
    LL    UL
 19517 45849
> t1 <-  predict(motor.wei, data.frame(temp=130),
                 type = "uquantile", p = 0.1, se = T)
> exp(c(LL=t1$fit - 2*t1$se, UL=t1$fit + 2*t1$se))
    LL    UL
 10258 24752
```

Nelson & Hahn worked with $z = 1000/(\texttt{temp} + 273.2)$. We leave the reader to try this; it gives slightly larger quantiles.

Function `censorReg`

S+ S-PLUS has a function `censorReg` for parametric survival analysis by Bill Meeker; this has a very substantial overlap with `survReg` but is more general in that it allows *truncation* as well as *censoring*. Either or both of censoring and truncation occur when subjects are only observed for part of the time axis. An observation T_i is right-censored if it is known only that $T_i > U_i$ for a censoring time U_i, and left-censored if it is known only that $T_i \leqslant L_i$. (Both left- and right-censoring can occur in a study, but not for the same individual.) Interval censoring is usually taken to refer to subjects known to have an event in $(L_i, U_i]$, but with the time of the event otherwise unknown. Truncation is similar but subtly different. For left and right truncation, subjects with events before L_i or after U_i are not included in the study, and interval truncation refers to both left and right truncation. (Notice the inconsistency with interval censoring.)

Confusingly, `censorReg` uses `"logexponential"` and `"lograyleigh"` for what are known to `survReg` as the `"exponential"` and `"rayleigh"` distributions and are accelerated-life models for those distributions.

Let us consider a simple example using `gehan`. We can fit a Weibull model by

```
> options(contrasts = c("contr.treatment", "contr.poly"))
> summary(censorReg(censor(time, cens) ~ treat, data = gehan))
    ....
Coefficients:
  Est. Std.Err. 95% LCL 95% UCL z-value  p-value
  3.52    0.252    3.02   4.009   13.96 2.61e-44
 -1.27    0.311   -1.88  -0.658   -4.08 4.51e-05

Extreme value distribution: Dispersion (scale) = 0.73219
Observations: 42 Total; 12 Censored
-2*Log-Likelihood: 213
```

which agrees with our results on page 363.

The potential advantages of `censorReg` come from its wider range of options. As noted previously, it allows truncation, by specifying a call to `censor` with a `truncation` argument. Distributions can be fitted with a *threshold*, that is, a parameter $\gamma > 0$ such that the failure-time model is fitted to $T - \gamma$ (and hence no failures can occur before time γ).

There is a `plot` method for `censorReg` that produces up to seven figures.

A `strata` argument in a `censorReg` model has a completely different meaning: it fits separate models at each level of the stratifying factor, unlike `survReg` which has common regression coefficients across strata.

13.3 Cox Proportional Hazards Model

Cox (1972) introduced a less parametric approach to proportional hazards. There is a baseline hazard function $h_0(t)$ that is modified multiplicatively by covariates

(including group indicators), so the hazard function for any individual case is

$$h(t) = h_0(t) \exp \boldsymbol{\beta}^T \boldsymbol{x}$$

and the interest is mainly in the proportional factors rather than the baseline hazard. Note that the cumulative hazards will also be proportional, so we can examine the hypothesis by plotting survivor curves for sub-groups on log scale. Later we allow the covariates to depend on time.

The parameter vector $\boldsymbol{\beta}$ is estimated by maximizing a *partial likelihood*. Suppose one death occurred at time t_j. Then conditional on this event the probability that case i died is

$$\frac{h_0(t) \exp \boldsymbol{\beta}^T \boldsymbol{x}_i}{\sum_l I(T_l \geqslant t) h_0(t) \exp \boldsymbol{\beta}^T \boldsymbol{x}_l} = \frac{\exp \boldsymbol{\beta}^T \boldsymbol{x}_i}{\sum_l I(T_l \geqslant t) \exp \boldsymbol{\beta}^T \boldsymbol{x}_l} \tag{13.11}$$

which does not depend on the baseline hazard. The partial likelihood for $\boldsymbol{\beta}$ is the product of such terms over all observed deaths, and usually contains most of the information about $\boldsymbol{\beta}$ (the remainder being in the observed times of death). However, we need a further condition on the censoring (Fleming and Harrington, 1991, pp. 138–9) that it is independent and *uninformative* for this to be so; the latter means that the likelihood for censored observations in $[t, t + \Delta t)$ does not depend on $\boldsymbol{\beta}$.

The correct treatment of ties causes conceptual difficulties as they are an event of probability zero for continuous distributions. Formally (13.11) may be corrected to include all possible combinations of deaths. As this increases the computational load, it is common to employ the Breslow approximation[4] in which each death is always considered to precede all other events at that time. Let $\tau_i = I(T_i \geqslant t) \exp \boldsymbol{\beta}^T \boldsymbol{x}_i$, and suppose there are d deaths out of m possible events at time t. Breslow's approximation uses the term

$$\prod_{i=1}^{d} \frac{\tau_i}{\sum_1^m \tau_j}$$

in the partial likelihood at time t. Other options are Efron's approximation

$$\prod_{i=1}^{d} \frac{\tau_i}{\sum_1^m \tau_j - \frac{i}{d} \sum_1^d \tau_j}$$

and the 'exact' partial likelihood

$$\prod_{i=1}^{d} \tau_i \Big/ \sum \prod_{k=1}^{d} \tau_{j_k}$$

where the sum is over subsets of $1, \ldots, m$ of size d. One of these terms is selected by the `method` argument of the function `coxph`, with default `efron`.

[4] First proposed by Peto (1972).

The baseline cumulative hazard $H_0(t)$ is estimated by rescaling the contributions to the number at risk by $\exp \widehat{\boldsymbol{\beta}}^T \boldsymbol{x}$ in (13.3). Thus in that formula $r(t) = \sum I(T_i \geqslant t) \exp \widehat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$.

The Cox model is easily extended to allow different baseline hazard functions for different groups, and this is automatically done if they are declared as `strata`. For our leukaemia example we have:

```
> leuk.cox <- coxph(Surv(time) ~ ag + log(wbc), data = leuk)
> summary(leuk.cox)
    ....
           coef exp(coef) se(coef)     z      p
      ag -1.069     0.343    0.429 -2.49 0.0130
log(wbc)  0.368     1.444    0.136  2.70 0.0069

         exp(coef) exp(-coef) lower .95 upper .95
      ag     0.343      2.913     0.148     0.796
log(wbc)     1.444      0.692     1.106     1.886

Rsquare= 0.377   (max possible= 0.994 )
Likelihood ratio test= 15.6  on 2 df,   p=0.000401
Wald test            = 15.1  on 2 df,   p=0.000537
Score (logrank) test = 16.5  on 2 df,   p=0.000263

> update(leuk.cox, ~ . -ag)
    ....
Likelihood ratio test=9.19  on 1 df, p=0.00243  n= 33

> (leuk.coxs <- coxph(Surv(time) ~ strata(ag) + log(wbc),
                      data = leuk))
    ....
          coef exp(coef) se(coef)    z      p
log(wbc) 0.391      1.48    0.143 2.74 0.0062
    ....
Likelihood ratio test=7.78  on 1 df, p=0.00529  n= 33

> (leuk.coxs1 <- update(leuk.coxs, . ~ . + ag:log(wbc)))
    ....
             coef exp(coef) se(coef)     z    p
   log(wbc) 0.183      1.20    0.188 0.978 0.33
ag:log(wbc) 0.456      1.58    0.285 1.598 0.11
    ....
> plot(survfit(Surv(time) ~ ag), lty = 2:3, log = T)
> lines(survfit(leuk.coxs), lty = 2:3, lwd = 3)
> legend(80, 0.8, c("ag absent", "ag present"), lty = 2:3)
```

The 'likelihood ratio test' is actually based on (log) partial likelihoods, not the full likelihood, but has similar asymptotic properties. The tests show that there is a significant effect of `wbc` on survival, but also that there is a significant difference between the two `ag` groups (although as Figure 13.4 shows, this is less than before adjustment for the effect of `wbc`).

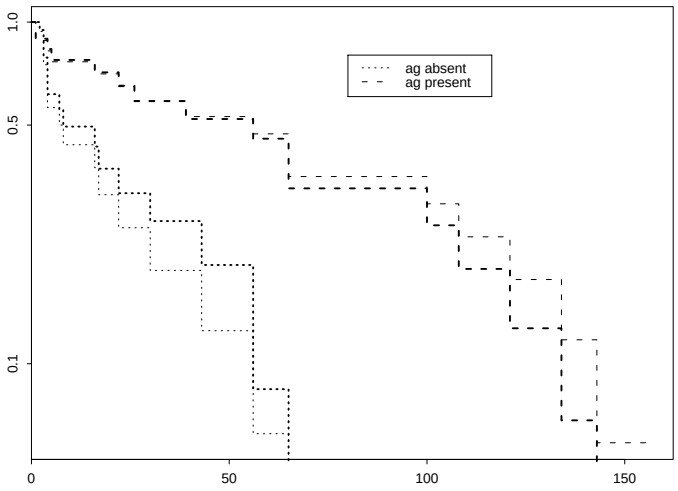

Figure 13.4: Log-survivor curves for the `leuk` dataset. The thick lines are from a Cox model with two strata, the thin lines Kaplan–Meier estimates that ignore the blood counts.

Note how `survfit` can take the result of a fit of a proportional hazard model. In the first fit the hazards in the two groups differ only by a factor whereas later they are allowed to have separate baseline hazards (which look very close to proportional). There is marginal evidence for a difference in slope within the two strata. Note how straight the log-survivor functions are in Figure 13.4, confirming the good fit of the exponential model for these data. The Kaplan–Meier survivor curves refer to the populations; those from the `coxph` fit refer to a patient in the stratum with an average `log(wbc)` for the whole dataset. This example shows why it is inappropriate just to test (using `survdiff`) the difference between the two groups; part of the difference is attributable to the lower `wbc` in the `ag` absent group.

The test statistics refer to the whole set of covariates. The likelihood ratio test statistic is the change in deviance on fitting the covariates over just the baseline hazard (by strata); the score test is the expansion at the baseline, and so does not need the parameters to be estimated (although this has been done). The R^2 measure quoted by `summary.coxph` is taken from Nagelkerke (1991).

The general proportional hazards model gives estimated (non-intercept) coefficients $\widehat{\boldsymbol{\beta}} = (-1.07, 0.37)^T$, compared to the Weibull fit of $(-0.98, 0.30)^T$ (on page 362). The log-logistic had coefficients $(-1.16, 0.61)^T$ which under the approximations of Solomon (1984) would be scaled by $\tau/2$ to give $(-0.79, 0.42)^T$ for a Cox proportional-hazards fit if the log-logistic regression model (an accelerated-life model) were the true model.

We next consider the Gehan data. We saw before that the pairing has a negligible effect for the exponential model. Here the effect is a little larger, with $P \approx 8\%$. The Gehan data have a large number of (mainly pairwise) ties, so we use the 'exact' partial likelihood.

```
> coxph(Surv(time, cens) ~ treat, data = gehan, method = "exact")
      coef exp(coef) se(coef)    z       p
treat 1.63      5.09    0.433 3.76 0.00017
```

```
Likelihood ratio test=16.2  on 1 df, p=5.54e-05  n= 42

# The next fit is slow
> coxph(Surv(time, cens) ~ treat + factor(pair), data = gehan,
        method = "exact")
    ....
Likelihood ratio test=45.5  on 21 df, p=0.00148  n= 42
    ....
> 1 - pchisq(45.5 - 16.2, 20)
[1] 0.082018
```

Finally we consider the `motors` data. The exact fit is much the slowest, as it has large groups of ties.

```
> (motor.cox <- coxph(Surv(time, cens) ~ temp, motors))
    ....
       coef exp(coef) se(coef)    z       p
temp 0.0918       1.1   0.0274 3.36 0.00079
    ....
> coxph(Surv(time, cens) ~ temp, motors, method = "breslow")
    ....
       coef exp(coef) se(coef)   z       p
temp 0.0905      1.09   0.0274 3.3 0.00098
    ....
> coxph(Surv(time, cens) ~ temp, motors, method = "exact")
    ....
       coef exp(coef) se(coef)    z       p
temp 0.0947       1.1   0.0274 3.45 0.00056
    ....
> plot( survfit(motor.cox, newdata = data.frame(temp=200),
                conf.type = "log-log") )
> summary( survfit(motor.cox, newdata = data.frame(temp=130)) )
time n.risk n.event survival  std.err lower 95% CI upper 95% CI
 408     40       4    1.000 0.000254        0.999            1
 504     36       3    1.000 0.000499        0.999            1
1344     28       2    0.999 0.001910        0.995            1
1440     26       1    0.998 0.002698        0.993            1
1764     20       1    0.996 0.005327        0.986            1
2772     19       1    0.994 0.007922        0.978            1
3444     18       1    0.991 0.010676        0.971            1
3542     17       1    0.988 0.013670        0.962            1
3780     16       1    0.985 0.016980        0.952            1
4860     15       1    0.981 0.020697        0.941            1
5196     14       1    0.977 0.024947        0.929            1
```

The function `survfit` has a special method for `coxph` objects that plots the mean and confidence interval of the survivor curve for an average individual (with average values of the covariates). As we see, this can be overridden by giving new data, as shown in Figure 13.5. The non-parametric method is unable to extrapolate to $130°C$ as none of the test examples survived long enough to estimate the baseline hazard beyond the last failure at 5 196 hours.

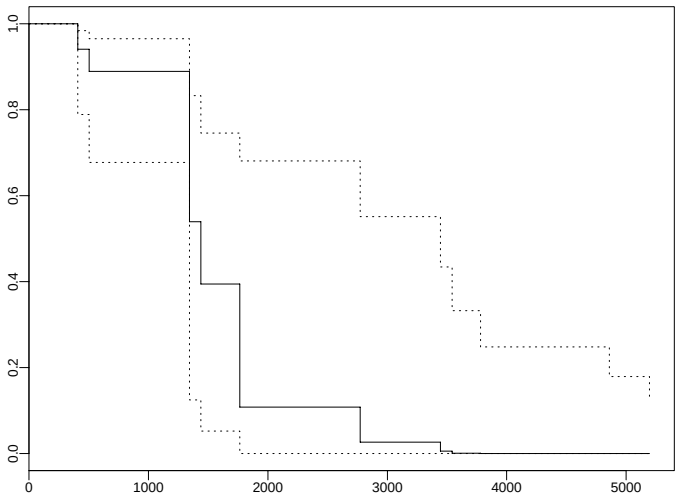

Figure 13.5: The survivor curve for a motor at $200^\circ\mathrm{C}$ estimated from a Cox proportional hazards model (solid line) with pointwise 95% confidence intervals (dotted lines).

Residuals

The concept of a residual is a difficult one for binary data, especially as here the event may not be observed because of censoring. A straightforward possibility is to take

$$r_i = \delta_i - \widehat{H}(t_i)$$

which is known as the *martingale residual* after a derivation from the mathematical theory given by Fleming and Harrington (1991, §4.5). They show that it is appropriate for checking the functional form of the proportional hazards model, for if

$$h(t) = h_0(t)\phi(x^*)\exp \boldsymbol{\beta}^T\boldsymbol{x}$$

for an (unknown) function of a covariate x^* then

$$E[R \mid X^*] \approx [\phi(X^*) - \overline{\phi}]\sum \delta_i/n$$

and this can be estimated by smoothing a plot of the martingale residuals versus x^*, for example, using `lowess` or the function `scatter.smooth` based on `loess`. (The term $\overline{\phi}$ is a complexly weighted mean.) The covariate x^* can be one not included in the model, or one of the terms to check for non-linear effects.

The martingale residuals are the default output of `residuals` on a `coxph` fit.

The martingale residuals can have a very skewed distribution, as their maximum value is 1, but they can be arbitrarily negative. The *deviance residuals* are a transformation

$$\text{sign}\,(r_i)\sqrt{2[-r_i - \delta_i \log(\delta_i - r_i)]}$$

which reduces the skewness, and for a parametric survival model when squared and summed give (approximately) the deviance. Deviance residuals are best used in plots that will indicate cases not fitted well by the model.

The *Schoenfeld residuals* (Schoenfeld, 1982) are defined at death times as $\boldsymbol{x}_i - \overline{\boldsymbol{x}}(t_i)$ where $\overline{\boldsymbol{x}}(s)$ is the mean weighted by $\exp \widehat{\boldsymbol{\beta}}^T\boldsymbol{x}$ of the $\boldsymbol{x}$ over only the cases still in the risk set at time s. These residuals form a matrix with one row for each case that died and a column for each covariate. The scaled Schoenfeld residuals (`type = "scaledsch"`) are the I^{-1} matrix multiplying the Schoenfeld residuals, where I is the (partial) information matrix at the fitted parameters in the Cox model.

The *score residuals* are the terms of efficient score for the partial likelihood, this being a sum over cases of

$$L_i = \big[\boldsymbol{x}_i - \overline{\boldsymbol{x}}(t_i)\big]\delta_i - \int_0^{t_i} \big[\boldsymbol{x}_i(s) - \overline{\boldsymbol{x}}(s)\big]\hat{h}(s)\,\mathrm{d}s$$

Thus the score residuals form an $n \times p$ matrix. They can be used to examine leverage of individual cases by computing (approximately) the change in $\widehat{\boldsymbol{\beta}}$ if the observation were dropped; `type = "dfbeta"` gives this, whereas `type = "dfbetas"` scales by the standard errors for the components of $\widehat{\boldsymbol{\beta}}$.

Tests of proportionality of hazards

Once a type of departure from the base model is discovered or suspected, the proportional hazards formulation is usually flexible enough to allow an extended model to be formulated and the significance of the departure tested within the extended model. Nevertheless, some approximations can be useful, and are provided by the function `cox.zph` for departures of the type

$$\boldsymbol{\beta}(t) = \boldsymbol{\beta} + \boldsymbol{\theta} g(t)$$

for some postulated smooth function g. Grambsch and Therneau (1994) show that the scaled Schoenfeld residuals for case i have, approximately, mean $g(t_i)\boldsymbol{\theta}$ and a computable variance matrix.

The function `cox.zph` has both `print` and `plot` methods. The printed output gives an estimate of the correlation between $g(t_i)$ and the scaled Schoenfeld residuals and a chi-squared test of $\boldsymbol{\theta} = 0$ for each covariate, and an overall chi-squared test. The plot method gives a plot for each covariate, of the scaled Schoenfeld residuals against $g(t)$ with a spline smooth and pointwise confidence bands for the smooth. (Figure 13.8 on page 375 is an example.)

The function g has to be specified. The default in `cox.zph` is $1 - \widehat{S}(t)$ for the Kaplan–Meier estimator, with options for the ranks of the death times, $g \equiv 1$ and $g = \log$ as well as a user-specified function. (The x-axis of the plots is labelled by the death times, not $\big\{g(t_i)\big\}$.)

13.4 Further Examples

VA lung cancer data

S-PLUS supplies[5] the dataset `cancer.vet` on a Veterans Administration lung cancer trial used by Kalbfleisch and Prentice (1980), but as it has no on-line help,

[5] For R it is supplied in package `MASS`.

it is not obvious what it is! It is a matrix of 137 cases with right-censored survival time and the covariates

treatment	standard or test
celltype	one of four cell types
Karnofsky score	of performance on scale 0–100, with high values for relatively well patients
diagnosis	time since diagnosis in months at entry to trial
age	in years
therapy	logical for prior therapy

As there are several covariates, we use the Cox model to establish baseline hazards.

```
> # R: data(VA) # is all that is required.
> # S: VA.temp <- as.data.frame(cancer.vet)
> # S: dimnames(VA.temp)[[2]] <- c("treat", "cell", "stime",
     "status", "Karn", "diag.time","age","therapy")
> # S: attach(VA.temp)
> # S: VA <- data.frame(stime, status, treat = factor(treat), age,
     Karn, diag.time, cell = factor(cell), prior = factor(therapy))
> # S: detach(VA.temp)
> (VA.cox <- coxph(Surv(stime, status) ~ treat + age  + Karn +
                   diag.time + cell + prior, data = VA))
               coef exp(coef) se(coef)        z       p
    treat  2.95e-01     1.343  0.20755  1.41945 1.6e-01
      age -8.71e-03     0.991  0.00930 -0.93612 3.5e-01
     Karn -3.28e-02     0.968  0.00551 -5.95801 2.6e-09
diag.time  8.18e-05     1.000  0.00914  0.00895 9.9e-01
    cell2  8.62e-01     2.367  0.27528  3.12970 1.7e-03
    cell3  1.20e+00     3.307  0.30092  3.97474 7.0e-05
    cell4  4.01e-01     1.494  0.28269  1.41955 1.6e-01
    prior  7.16e-02     1.074  0.23231  0.30817 7.6e-01

Likelihood ratio test=62.1  on 8 df, p=1.8e-10  n= 137

> (VA.coxs <- coxph(Surv(stime, status) ~ treat + age + Karn +
      diag.time + strata(cell) + prior, data = VA))
              coef exp(coef) se(coef)      z       p
    treat  0.28590     1.331  0.21001  1.361 1.7e-01
      age -0.01182     0.988  0.00985 -1.201 2.3e-01
     Karn -0.03826     0.962  0.00593 -6.450 1.1e-10
diag.time -0.00344     0.997  0.00907 -0.379 7.0e-01
    prior  0.16907     1.184  0.23567  0.717 4.7e-01

Likelihood ratio test=44.3  on 5 df, p=2.04e-08  n= 137

> plot(survfit(VA.coxs), log = T, lty = 1:4, col = 2:5)
> legend(locator(1), c("squamous", "small", "adeno", "large"),
     lty = 1:4, col = 2:5)
> plot(survfit(VA.coxs), fun = "cloglog", lty = 1:4, col = 2:5)
```

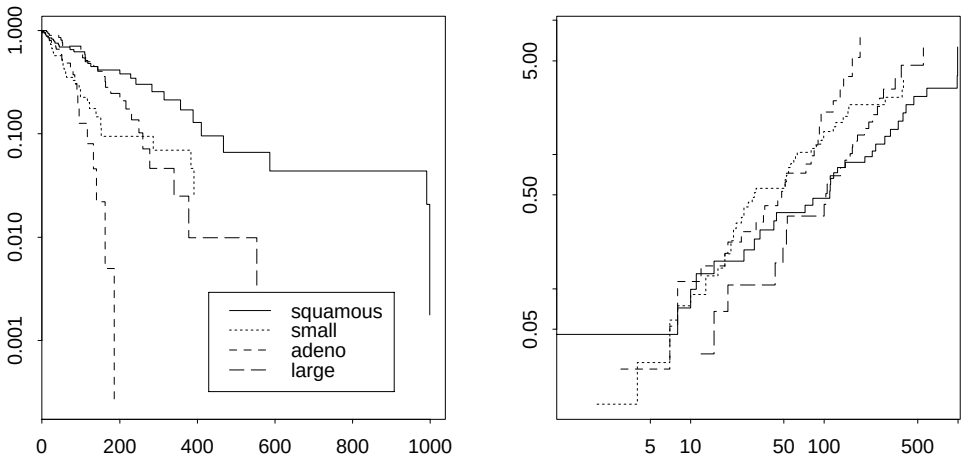

Figure 13.6: Cumulative hazard functions for the cell types in the VA lung cancer trial. The left-hand plot is labelled by survival probability on log scale. The right-hand plot is on log-log scale.

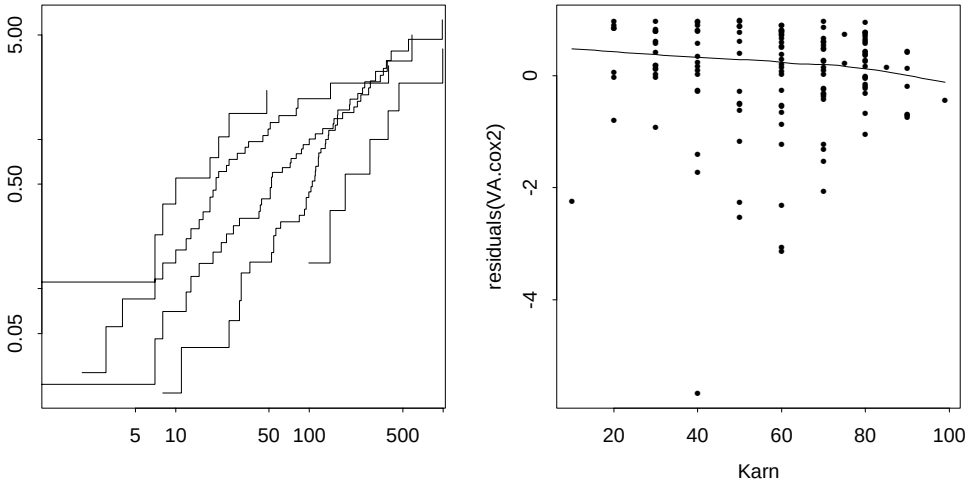

Figure 13.7: Diagnostic plots for the Karnofsky score in the VA lung cancer trial. Left: log-log cumulative hazard plot for five groups. Right: martingale residuals versus Karnofsky score, with a smoothed fit.

```
> cKarn <- factor(cut(VA$Karn, 5))
> VA.cox1 <- coxph(Surv(stime, status) ~ strata(cKarn) + cell,
                   data = VA)
> plot(survfit(VA.cox1), fun="cloglog")
> VA.cox2 <- coxph(Surv(stime, status) ~ Karn + strata(cell),
                   data = VA)
> # R: library(modreg)
> scatter.smooth(VA$Karn, residuals(VA.cox2))
```

Figures 13.6 and 13.7 show some support for proportional hazards among the cell types (except perhaps squamous), and suggest a Weibull or even exponential distribution.

```
> VA.wei <- survReg(Surv(stime, status) ~ treat + age + Karn +
     diag.time + cell + prior, data = VA)
```

```
> summary(VA.wei, cor = F)
    ....
               Value Std. Error       z        p
(Intercept)  3.262014    0.66253  4.9236 8.50e-07
      treat -0.228523    0.18684 -1.2231 2.21e-01
        age  0.006099    0.00855  0.7131 4.76e-01
       Karn  0.030068    0.00483  6.2281 4.72e-10
  diag.time -0.000469    0.00836 -0.0561 9.55e-01
      cell2 -0.826185    0.24631 -3.3542 7.96e-04
      cell3 -1.132725    0.25760 -4.3973 1.10e-05
      cell4 -0.397681    0.25475 -1.5611 1.19e-01
      prior -0.043898    0.21228 -0.2068 8.36e-01
 Log(scale) -0.074599    0.06631 -1.1250 2.61e-01

Scale= 0.928

Weibull distribution
Loglik(model)= -715.6   Loglik(intercept only)= -748.1
        Chisq= 65.08 on 8 degrees of freedom, p= 4.7e-11

> VA.exp <- survReg(Surv(stime, status) ~ Karn + cell,
                    data = VA, dist = "exponential")
> summary(VA.exp, cor = F)
              Value Std. Error     z        p
(Intercept)  3.4222    0.35463  9.65 4.92e-22
       Karn  0.0297    0.00486  6.11 9.97e-10
      cell2 -0.7102    0.24061 -2.95 3.16e-03
      cell3 -1.0933    0.26863 -4.07 4.70e-05
      cell4 -0.3113    0.26635 -1.17 2.43e-01

Scale fixed at 1

Exponential distribution
Loglik(model)= -717   Loglik(intercept only)= -751.2
        Chisq= 68.5 on 4 degrees of freedom, p= 4.7e-14
```

Note that `scale` does not differ significantly from one, so an exponential distribution is an appropriate summary.

```
> cox.zph(VA.coxs)
              rho  chisq       p
    treat -0.0607  0.545 0.46024
      age  0.1734  4.634 0.03134
     Karn  0.2568  9.146 0.00249
diag.time  0.1542  2.891 0.08909
    prior -0.1574  3.476 0.06226
   GLOBAL      NA 13.488 0.01921
> par(mfrow = c(3, 2)); plot(cox.zph(VA.coxs))
```

Closer investigation does show some suggestion of time-varying coefficients in the Cox model. The plot is Figure 13.8. Note that some coefficients that are not

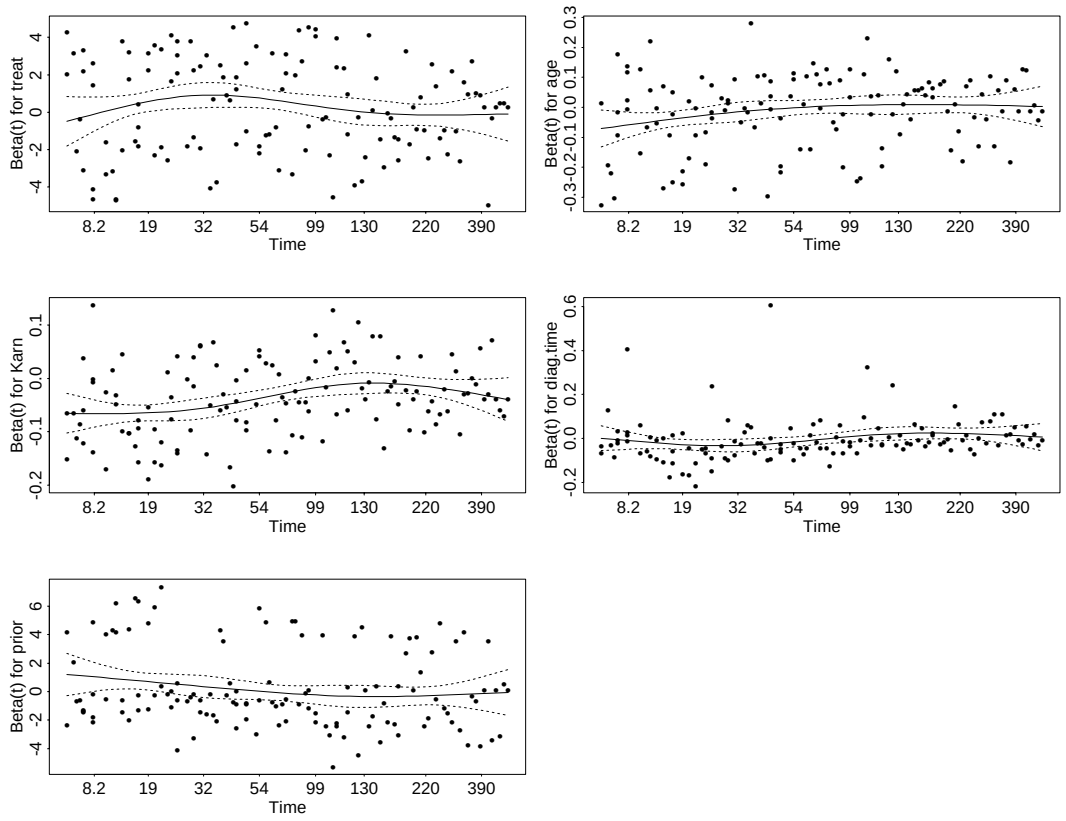

Figure 13.8: Diagnostics plots from `cox.zph` of the constancy of the coefficients in the proportional hazards model `VA.coxs`. Each plot is of a component of the Schoenfeld residual against a non-linear scale of time. A spline smoother is shown, together with ± 2 standard deviations.

significant in the basic model show evidence of varying with time. This suggests that the model with just `Karn` and `cell` may be too simple, and that we need to consider interactions. We automate the search of interactions using `stepAIC`, which has methods for both `coxph` and `survReg` fits. With hindsight, we centre the data.

```
> VA$Karnc <- VA$Karn - 50
> VA.coxc <- update(VA.cox, ~ . - Karn + Karnc)
> VA.cox2 <- stepAIC(VA.coxc, ~ .^2)
> VA.cox2$anova
Initial Model:
Surv(stime, status) ~ treat + age + diag.time + cell + prior +
        Karnc

Final Model:
Surv(stime, status) ~ treat + diag.time + cell + prior + Karnc +
        prior:Karnc + diag.time:cell + treat:prior + treat:Karnc

              Step Df Deviance Resid. Df Resid. Dev    AIC
1                                    129     948.79 964.79
2    + prior:Karnc  1    9.013       128     939.78 957.78
```

```
3 + diag.time:cell  3   11.272       125     928.51 952.51
4            - age  1    0.415       126     928.92 950.92
5    + treat:prior  1    2.303       125     926.62 950.62
6    + treat:Karnc  1    2.904       124     923.72 949.72
```

(The 'deviances' here are minus twice log partial likelihoods.) Applying `stepAIC` to `VA.wei` leads to the same sequence of steps. As the variables `diag.time` and `Karn` are not factors, this will be easier to interpret using nesting:

```
> (VA.cox3 <- update(VA.cox2, ~ treat/Karnc + prior*Karnc
    + treat:prior + cell/diag.time))
                    coef exp(coef) se(coef)      z       p
           treat  0.8065     2.240  0.27081  2.978 2.9e-03
           prior  0.9191     2.507  0.31568  2.912 3.6e-03
           Karnc -0.0107     0.989  0.00949 -1.129 2.6e-01
           cell2  1.7068     5.511  0.37233  4.584 4.6e-06
           cell3  1.5633     4.775  0.44205  3.536 4.1e-04
           cell4  0.7476     2.112  0.48136  1.553 1.2e-01
Karnc %in% treat -0.0187     0.981  0.01101 -1.695 9.0e-02
     prior:Karnc -0.0481     0.953  0.01281 -3.752 1.8e-04
     treat:prior -0.7264     0.484  0.41833 -1.736 8.3e-02
  cell1diag.time  0.0532     1.055  0.01595  3.333 8.6e-04
  cell2diag.time -0.0245     0.976  0.01293 -1.896 5.8e-02
  cell3diag.time  0.0161     1.016  0.04137  0.388 7.0e-01
  cell4diag.time  0.0150     1.015  0.04033  0.373 7.1e-01
```

Thus the hazard increases with time since diagnosis in squamous cells, only, and the effect of the Karnofsky score is only pronounced in the group with prior therapy. We tried replacing `diag.time` with a polynomial, with negligible benefit. Using `cox.zph` shows a very significant change with time in the coefficients of the `treat*Karn` interaction.

```
> cox.zph(VA.cox3)
                      rho    chisq        p
           treat  0.18012  6.10371 0.013490
           prior  0.07197  0.76091 0.383044
           Karnc  0.27220 14.46103 0.000143
           cell2  0.09053  1.31766 0.251013
           cell3  0.06247  0.54793 0.459164
           cell4  0.00528  0.00343 0.953318
Karnc %in% treat -0.20606  7.80427 0.005212
     prior:Karnc -0.04017  0.26806 0.604637
     treat:prior -0.13061  2.33270 0.126682
  cell1diag.time  0.11067  1.62464 0.202446
  cell2diag.time -0.01680  0.04414 0.833596
  cell3diag.time  0.09713  1.10082 0.294086
  cell4diag.time  0.16912  3.16738 0.075123
          GLOBAL       NA 25.52734 0.019661

> par(mfrow = c(2, 2))
> plot(cox.zph(VA.cox3), var = c(1, 3, 7)) ## not shown
```

Stanford heart transplants

This set of data is analysed by Kalbfleisch and Prentice (1980, §5.5.3). (The data given in Kalbfleisch & Prentice are rounded, but the full data are supplied as data frame `heart`.) It is on survival from early heart transplant operations at Stanford. The new feature is that patients may change treatment during the study, moving from the control group to the treatment group at transplantation, so some of the covariates such as waiting time for a transplant are time-dependent (in the simplest possible way). Patients who received a transplant are treated as two cases, before and after the operation, so cases in the transplant group are in general both right-censored and left-truncated. This is handled by `Surv` by supplying entry and exit times. For example, patient 4 has the rows

```
start   stop event      age        year   surgery transplant
  0.0   36.0   0  -7.73716632  0.49007529       0          0
 36.0   39.0   1  -7.73716632  0.49007529       0          1
```

which show that he waited 36 days for a transplant and then died after 3 days. The proportional hazards model is fitted from this set of cases, but some summaries need to take account of the splitting of patients.

The covariates are age (in years minus 48), year (after 1 October 1967) and an indicator for previous surgery. Rather than use the six models considered by Kalbfleisch & Prentice, we do our own model selection.

```
> coxph(Surv(start, stop, event) ~ transplant*
     (age + surgery + year), data = heart)
    ....
Likelihood ratio test=18.9  on 7 df, p=0.00852  n= 172
> coxph(Surv(start, stop, event) ~ transplant*(age + year) +
     surgery, data = heart)
    ....
Likelihood ratio test=18.4  on 6 df, p=0.0053  n= 172
> (stan <- coxph(Surv(start, stop, event) ~ transplant*year +
     age + surgery, data = heart))
    ....
                 coef exp(coef) se(coef)     z     p
     transplant -0.6213     0.537   0.5311 -1.17 0.240
           year -0.2526     0.777   0.1049 -2.41 0.016
            age  0.0299     1.030   0.0137  2.18 0.029
        surgery -0.6641     0.515   0.3681 -1.80 0.071
transplant:year  0.1974     1.218   0.1395  1.42 0.160

Likelihood ratio test=17.1  on 5 df, p=0.00424  n= 172

> stan1 <- coxph(Surv(start, stop, event) ~ strata(transplant) +
     year + year:transplant + age + surgery, heart)
> plot(survfit(stan1), conf.int = T, log = T, lty = c(1, 3),
       col = 2:3)
> legend(locator(1), c("before", "after"), lty = c(1, 3),
         col = 2:3)
```

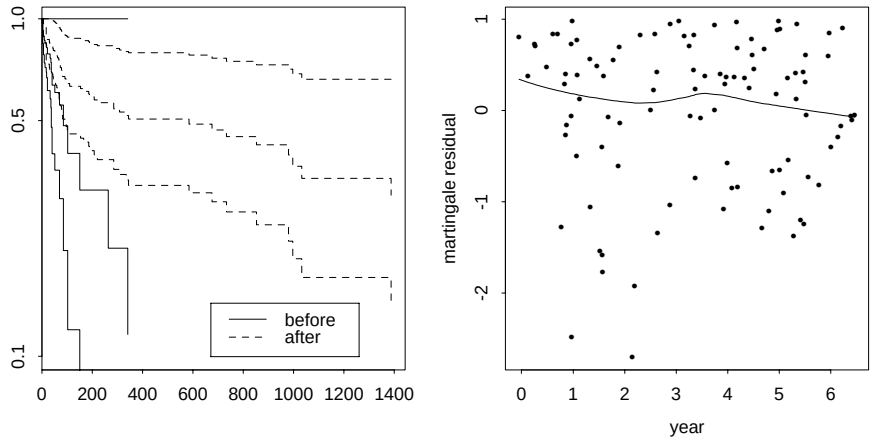

Figure 13.9: Plots for the Stanford heart transplant study. Left: log survivor curves and confidence limits for the two groups. Right: martingale residuals against calendar time.

```
> attach(heart)
> plot(year[transplant==0], residuals(stan1, collapse = id),
      xlab = "year", ylab = "martingale residual")
> lines(lowess(year[transplant == 0],
               residuals(stan1, collapse = id)))
> sresid <- resid(stan1, type = "dfbeta", collapse = id)
> detach()
> -100 * sresid %*% diag(1/stan1$coef)
```

This analysis suggests that survival rates over the study improved *prior* to transplantation, which Kalbfleisch & Prentice suggest could be due to changes in recruitment. The diagnostic plots of Figure 13.9 show nothing amiss. The `collapse` argument is needed as those patients who received transplants are treated as two cases, and we need the residual per patient.

Now consider predicting the survival of future patient aged 50 on 1 October 1971 with prior surgery, transplanted after six months.

```
# Survivor curve for the "average" subject
> summary(survfit(stan))
#  follow-up for two years
> stan2 <- data.frame(start = c(0, 183), stop= c(183, 2*365),
     event = c(0, 0), year = c(4, 4), age = c(50, 50) - 48,
     surgery = c(1, 1), transplant = c(0, 1))
> summary(survfit(stan, stan2, individual = T,
                  conf.type = "log-log"))
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    ....
  165     43       1    0.654 0.11509        0.384        0.828
  186     41       1    0.643 0.11602        0.374        0.820
  188     40       1    0.632 0.11697        0.364        0.812
  207     39       1    0.621 0.11790        0.353        0.804
  219     38       1    0.610 0.11885        0.343        0.796
  263     37       1    0.599 0.11978        0.332        0.788
```

```
285    35    2    0.575 0.11524    0.325    0.762
308    33    1    0.564 0.11618    0.314    0.753
334    32    1    0.552 0.11712    0.302    0.744
340    31    1    0.540 0.11799    0.291    0.735
343    29    1    0.527 0.11883    0.279    0.725
584    21    1    0.511 0.12018    0.263    0.713
675    17    1    0.492 0.12171    0.245    0.699
```

The argument `individual = T` is needed to avoid averaging the two cases (which are the same individual).

Australian AIDS survival

The data on the survival of AIDS patients within Australia are of unusually high quality within that field, and jointly with Dr Patty Solomon we have studied survival up to 1992.[6] There are a large number of difficulties in defining survival from AIDS (acquired immunodeficiency syndrome), in part because as a syndrome its diagnosis is not clear-cut and has almost certainly changed with time. (To avoid any possible confusion, we are studying survival from AIDS and not the HIV infection which is generally accepted as the cause of AIDS.)

The major covariates available were the reported transmission category, and the state or territory within Australia. The AIDS epidemic had started in New South Wales and then spread, so the states have different profiles of cases in calendar time. A factor that was expected to be important in survival is the widespread availability of zidovudine (AZT) to AIDS patients from mid-1987 which has enhanced survival, and the use of zidovudine for HIV-infected patients from mid-1990, which it was thought might delay the onset of AIDS without necessarily postponing death further.

The transmission categories were:

hs	male homosexual or bisexual contact
hsid	as **hs** and also intravenous drug user
id	female or heterosexual male intravenous drug user
het	heterosexual contact
haem	haemophilia or coagulation disorder
blood	receipt of blood, blood components or tissue
mother	mother with or at risk of HIV infection
other	other or unknown

The data file gave data on all patients whose AIDS status was diagnosed prior to January 1992, with their status then. Since there is a delay in notification of death, some deaths in late 1991 would not have been reported and we adjusted the endpoint of the study to 1 July 1991. A total of 2 843 patients were included, of whom about 1 770 had died by the end date. The file contained an ID number, the dates of first diagnosis, birth and death (if applicable), as well as the state and the coded transmission category. We combined the states `ACT` and `NSW` (as

[6] We are grateful to the Australian National Centre in HIV Epidemiology and Clinical Research for making these data available to us.

Australian Capital Territory is a small enclave within New South Wales), and to maintain confidentiality the dates have been jittered and the smallest states combined. Only the transformed file `Aids2` is included in our library.

As there are a number of patients who are diagnosed at (strictly, after) death, there are a number of zero survivals. The software used to have problems with these, so all deaths were shifted by 0.9 days to occur after other events the same day. To transform `Aids2` to a form suitable for time-dependent-covariate analysis we used

```
time.depend.covar <- function(data) {
  id <- row.names(data); n <- length(id)
  events <- c(0, 10043, 11139, 12053) # julian days
  crit1 <- matrix(events[1:3], n, 3 ,byrow = T)
  crit2 <- matrix(events[2:4], n, 3, byrow = T)
  diag <- matrix(data$diag,n,3); death <- matrix(data$death,n,3)
  incid <- (diag < crit2) & (death >= crit1); incid <- t(incid)
  indr <- col(incid)[incid]; indc <- row(incid)[incid]
  ind <- cbind(indr, indc); idno <- id[indr]
  state <- data$state[indr]; T.categ <- data$T.categ[indr]
  age <- data$age[indr]; sex <- data$sex[indr]
  late <- indc - 1
  start <- t(pmax(crit1 - diag, 0))[incid]
  stop <- t(pmin(crit2, death + 0.9) - diag)[incid]
  status <- matrix(as.numeric(data$status),n,3)-1 # 0/1
  status[death > crit2] <- 0; status <- status[ind]
  levels(state) <- c("NSW", "Other", "QLD", "VIC")
  levels(T.categ) <- c("hs", "hsid", "id", "het", "haem",
                       "blood", "mother", "other")
  levels(sex) <- c("F", "M")
  data.frame(idno, zid=factor(late), start, stop, status,
             state, T.categ, age, sex)
}
Aids3 <- time.depend.covar(Aids2)
```

The factor `zid` indicates whether the patient is likely to have received zidovudine at all, and if so whether it might have been administered during HIV infection.

Our analysis was based on a proportional hazards model that allowed a proportional change in hazard from 1 July 1987 to 30 June 1990 and another from 1 July 1990; the results show a halving of hazard from 1 July 1987 but a nonsignificant change in 1990.

```
> attach(Aids3)
> aids.cox <- coxph(Surv(start, stop, status)
      ~ zid + state + T.categ + sex + age, data = Aids3)
> summary(aids.cox)

                  coef exp(coef) se(coef)        z       p
        zid1 -0.69087     0.501  0.06578 -10.5034 0.0e+00
        zid2 -0.78274     0.457  0.07550 -10.3675 0.0e+00
  stateOther -0.07246     0.930  0.08964  -0.8083 4.2e-01
```

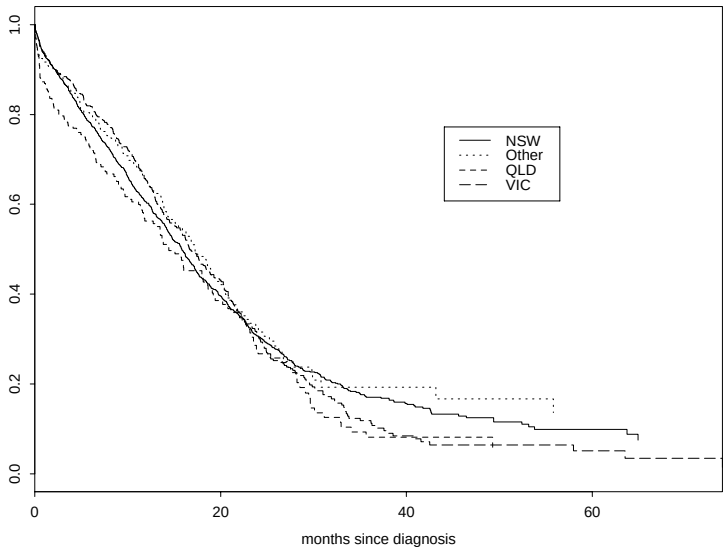

Figure 13.10: Survival of AIDS patients in Australia by state.

```
     stateQLD  0.18315     1.201  0.08752   2.0927 3.6e-02
     stateVIC  0.00464     1.005  0.06134   0.0756 9.4e-01
  T.categhsid -0.09937     0.905  0.15208  -0.6534 5.1e-01
    T.categid -0.37979     0.684  0.24613  -1.5431 1.2e-01
   T.categhet -0.66592     0.514  0.26457  -2.5170 1.2e-02
  T.categhaem  0.38113     1.464  0.18827   2.0243 4.3e-02
 T.categblood  0.16856     1.184  0.13763   1.2248 2.2e-01
T.categmother  0.44448     1.560  0.58901   0.7546 4.5e-01
 T.categother  0.13156     1.141  0.16380   0.8032 4.2e-01
          sex  0.02421     1.025  0.17557   0.1379 8.9e-01
          age  0.01374     1.014  0.00249   5.5060 3.7e-08
    ....
Likelihood ratio test= 185  on 14 df,   p=0
```

The effect of `sex` is nonsignificant, and so dropped in further analyses. There is no detected difference in survival during 1990.

Note that Queensland has a significantly elevated hazard relative to New South Wales (which has over 60% of the cases), and that the intravenous drug users have a longer survival, whereas those infected via blood or blood products have a shorter survival, relative to the first category who form 87% of the cases. We can use stratified Cox models to examine these effects (Figures 13.10 and 13.11).

```
> aids1.cox <- coxph(Surv(start, stop, status)
   ~ zid + strata(state) + T.categ + age, data = Aids3)
> (aids1.surv <- survfit(aids1.cox))
              n events mean se(mean) median 0.95LCL 0.95UCL
  state=NSW 1780   1116  639     17.6    481     450     509
state=Other  249    142  658     42.2    525     453     618
  state=QLD  226    149  519     33.5    439     360     568
  state=VIC  588    355  610     26.3    508     476     574
> plot(aids1.surv, mark.time = F, lty = 1:4, col = 2:5,
      xscale = 365.25/12, xlab = "months since diagnosis")
```

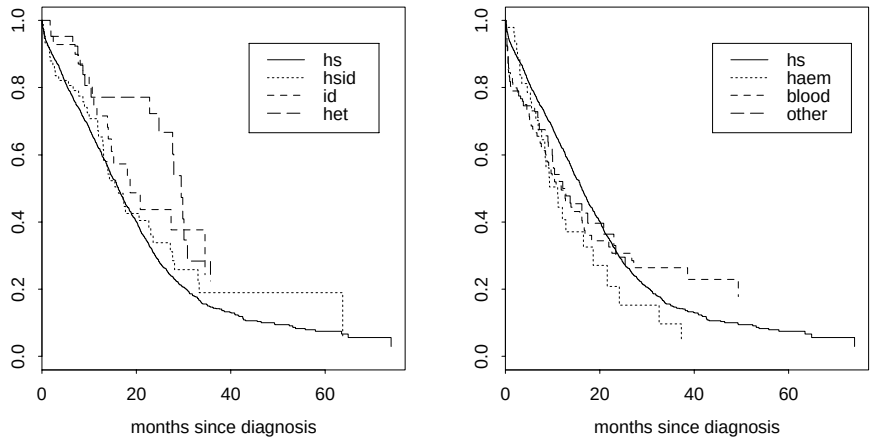

Figure 13.11: Survival of AIDS patients in Australia by transmission category.

```
> legend(locator(1), levels(state), lty = 1:4, col = 2:5)

> aids2.cox <- coxph(Surv(start, stop, status)
   ~ zid + state + strata(T.categ) + age, data = Aids3)
> (aids2.surv <- survfit(aids2.cox))
                 n events mean se(mean) median 0.95LCL 0.95UCL
    T.categ=hs 2465   1533  633     15.6    492   473.9     515
  T.categ=hsid   72     45  723     86.7    493   396.9     716
    T.categ=id   48     19  653     54.3    568   447.9      NA
   T.categ=het   40     17  775     57.3    897   842.9      NA
  T.categ=haem   46     29  431     53.9    337   252.9     657
 T.categ=blood   94     76  583     86.1    358   267.9     507
T.categ=mother    7      3  395     92.6    655    15.9      NA
 T.categ=other   70     40  421     40.7    369   300.9     712

> par(mfrow = c(1, 2))
> plot(aids2.surv[1:4], mark.time = F, lty = 1:4, col = 2:5,
   xscale = 365.25/12, xlab = "months since diagnosis")
> legend(locator(1), levels(T.categ)[1:4], lty = 1:4, col = 2:5)

> plot(aids2.surv[c(1, 5, 6, 8)], mark.time = F, lty = 1:4,
   col = 2:5, xscale = 365.25/12, xlab = "months since diagnosis")
> legend(locator(1), levels(T.categ)[c(1, 5, 6, 8)],
         lty = 1:4, col = 2:5)
```

We now consider the possible non-linear dependence of log-hazard on `age`. First we consider the martingale residual plot.

```
cases <- diff(c(0,idno)) != 0
aids.res <- residuals(aids.cox, collapse = idno)
scatter.smooth(age[cases], aids.res, xlab = "age",
               ylab = "martingale residual")
```

This shows a slight rise in residual with age over 60, but no obvious effect. The next step is to augment a linear term in age by a step function, with breaks chosen

from prior experience. We set the base level to be the 31–40 age group by using `relevel`, which re-orders the factor levels.

```
age2 <- cut(age, c(-1, 15, 30, 40, 50, 60, 100))
c.age <- factor(as.numeric(age2), labels = c("0-15", "16-30",
   "31-40", "41-50", "51-60", "61+"))
table(c.age)
 0-15 16-30 31-40 41-50 51-60 61+
   39  1022  1583   987   269  85
c.age <- relevel(c.age, "31-40")

summary(coxph(Surv(start, stop, status) ~ zid  + state
   + T.categ + age + c.age, data = Aids3))
    ....
               coef   exp(coef) se(coef)      z       p
    ....
       age  0.009218     1.009  0.00818   1.1266 0.2600
 c.age0-15  0.499093     1.647  0.36411   1.3707 0.1700
c.age16-30 -0.019631     0.981  0.09592  -0.2047 0.8400
c.age41-50 -0.004818     0.995  0.09714  -0.0496 0.9600
c.age51-60  0.198136     1.219  0.18199   1.0887 0.2800
  c.age61+  0.413690     1.512  0.30821   1.3422 0.1800
    ....
Likelihood ratio test= 193  on 18 df,   p=0
    ....
detach()
```

which is not a significant improvement in fit. Beyond this we could fit a smooth function of `age` via splines, but to save computational time we deferred this to the parametric analysis, which we now consider. From the survivor curves the obvious model is the Weibull. Since this is both a proportional hazards model and an accelerated-life model, we can include the effect of the introduction of zidovudine by assuming a doubling of survival after July 1987. With 'time' computed on this basis we find

```
make.aidsp <- function(){
   cutoff <- 10043
   btime <- pmin(cutoff, Aids2$death) - pmin(cutoff, Aids2$diag)
   atime <- pmax(cutoff, Aids2$death) - pmax(cutoff, Aids2$diag)
   survtime <- btime + 0.5*atime
   status <- as.numeric(Aids2$status)
   data.frame(survtime, status = status - 1, state = Aids2$state,
     T.categ = Aids2$T.categ, age = Aids2$age, sex = Aids2$sex)
}

Aidsp <- make.aidsp()
aids.wei <- survReg(Surv(survtime + 0.9, status) ~  state
                    + T.categ + sex + age, data = Aidsp)
summary(aids.wei, cor = F)
    ....
Coefficients:
```

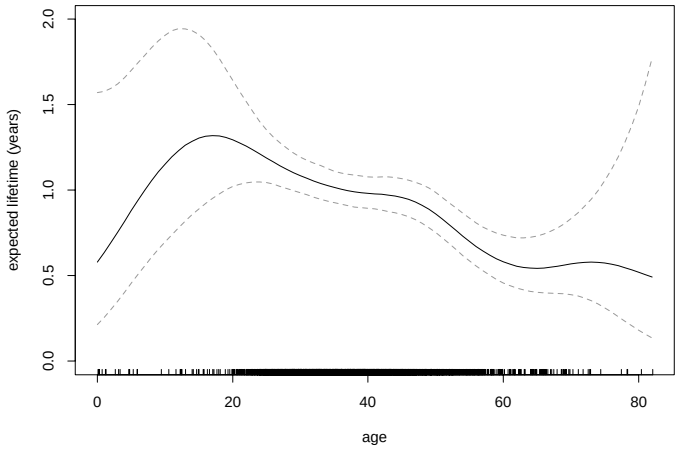

Figure 13.12: Predicted survival versus age of a `NSW hs` patient (solid line), with pointwise 95% confidence intervals (dashed lines) and a rug of all observed ages.

```
                Value Std. Error        z         p
  (Intercept)  6.41825     0.2098  30.5970 1.34e-205
   stateOther  0.09387     0.0931   1.0079  3.13e-01
     stateQLD -0.18213     0.0913  -1.9956  4.60e-02
     stateVIC -0.00750     0.0637  -0.1177  9.06e-01
  T.categhsid  0.09363     0.1582   0.5918  5.54e-01
    T.categid  0.40132     0.2552   1.5727  1.16e-01
   T.categhet  0.67689     0.2744   2.4667  1.36e-02
  T.categhaem -0.34090     0.1956  -1.7429  8.14e-02
 T.categblood -0.17336     0.1429  -1.2131  2.25e-01
T.categmother -0.40186     0.6123  -0.6563  5.12e-01
 T.categother -0.11279     0.1696  -0.6649  5.06e-01
          sex -0.00426     0.1827  -0.0233  9.81e-01
          age -0.01374     0.0026  -5.2862  1.25e-07
   Log(scale)  0.03969     0.0193   2.0572  3.97e-02

Scale= 1.04
```

Note that we continue to avoid zero survival. This shows good agreement with the parameters for the Cox model. The parameter α (the reciprocal of the scale) is close to one. For practical purposes the exponential is a good fit, and the parameters are little changed.

We also considered parametric non-linear functions of `age` by using a spline function. We use the P-splines of Eilers and Marx (1996) as this is implemented in both `survReg` and `coxph`; it can be seen as a convenient approximation to smoothing splines. For useful confidence intervals we include the constant term in the predictions, which are for a `NSW hs` patient. Note that for valid prediction with `pspline` the range of the new data must exactly match that of the old data.

```
> survReg(Surv(survtime + 0.9, status) ~ state + T.categ
   + age, data = Aidsp)
    ....
Scale= 1.0405
```

```
Loglik(model)= -12111   Loglik(intercept only)= -12140

> (aids.ps <- survReg(Surv(survtime + 0.9, status) ~  state
    + T.categ + pspline(age, df=6), data = Aidsp))
    ....
                               coef se(coef)     se2 Chisq   DF
            (Intercept)  4.83189 0.82449  0.60594 34.34 1.00
    ....
pspline(age, df = 6), lin -0.01362 0.00251  0.00251 29.45 1.00
pspline(age, df = 6), non                           9.82 5.04
                                 p
    ....
pspline(age, df = 6), lin 5.8e-08
pspline(age, df = 6), non 8.3e-02
    ....
> zz <- predict(aids.ps, data.frame(
    state = factor(rep("NSW", 83), levels = levels(Aidsp$state)),
    T.categ = factor(rep("hs", 83), levels = levels(Aidsp$T.categ)),
    age = 0:82), se = T, type = "linear")
> plot(0:82, exp(zz$fit)/365.25, type = "l", ylim = c(0, 2),
    xlab = "age", ylab = "expected lifetime (years)")
> lines(0:82, exp(zz$fit+1.96*zz$se.fit)/365.25, lty = 3, col = 2)
> lines(0:82, exp(zz$fit-1.96*zz$se.fit)/365.25, lty = 3, col = 2)
> rug(Aidsp$age + runif(length(Aidsp$age), -0.5, 0.5),
      ticksize = 0.015)
```

The results (Figure 13.12) suggest that a non-linear in age term is not worthwhile, although there are too few young people to be sure. We predict log-time to get confidence intervals on that scale.