M463: Survival and Event History Analysis

Contents

1 Introduction

1.1 Administration

Books:

Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1992). Statistical Models Based on Counting Processes.

Fleming, T.R. and Harrington, D.P. (1991). Counting Processes and Survival Analysis. Wiley.

Hosmer, D.W. and Lemeshow, S. (1999). Applied Survival Analysis. Wiley.

Hougaard, P. (2000). Analysis of Multivariate Survival Data. Springer.

Klein, J.P. and Moeschberger, M.L. (1997). Survival Analysis. Springer.

Therneau, T.M. and Grambsch, P.M. (2000). Modeling (sic) Survival Data: Extending the Cox Model. Springer.

Assessment:

20%	exercises
30%	project
50%	examination

1.2 Background

This is a second course in survival analysis. Hence familiarity with the basics is assumed, though Section 2 provides a quick recap of the methods which are most commonly used in the analysis of medical survival data: Kaplan-Meier, log rank, proportional hazards. The remainder of the course provides a selection from the huge variety of techniques which are available for more refined analysis in this important area together with a flavour of the underlying theory.

Standard notation (as opposed to counting process notation - see Section 4) is

- t_i $(i = 1, 2, \ldots, n)$: observation time
- δ_i: censoring/failure indicator - 0 if censored, 1 if failure
- x_i: p-vector of covariates

We will assume that the covariates x_i do not vary over time, though for most of the methods this is not a necessary assumption. Note that we use “failure” to denote the event of interest, even though this need not be death. It could be discharge from hospital, recurrence of a condition, an event such as seizure or MI, and so on.

Of course one unusual aspect of survival data is that usually we have incomplete information. This can occur for a variety of reasons, including the following.

- Right censoring. Actual failure times are not observed for some patients.
- Left censoring. Some start times may not be observed.

- Interval censoring. Failure (or start) is known only to lie within some range.

- Current status data - only know whether an event has or has not occurred at the study time.

- Truncation - screening some people so that investigator not aware of existence. Eg prevalent cohorts - those already having had the event are excluded.

A *Lexis diagram* can help with understanding the difference. Further explanation will be given in lectures.

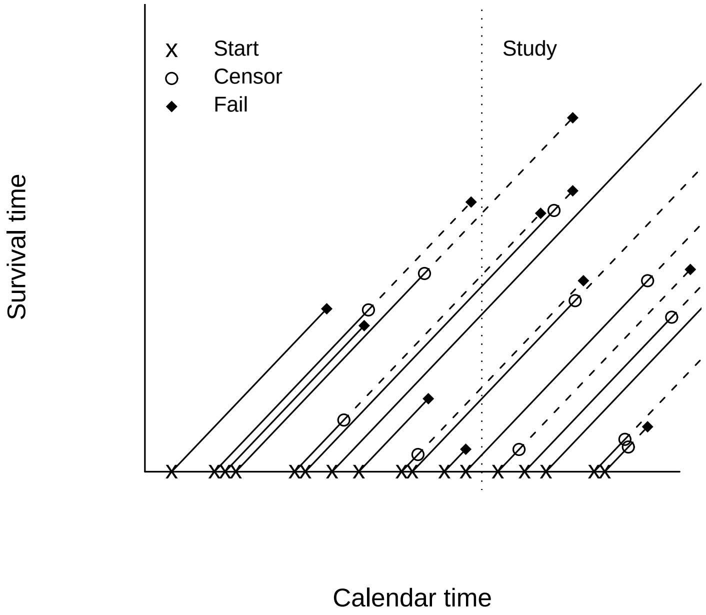


Figure 1: A Lexis diagram

For this course we assume *independent right censoring* only: extension to other censoring or truncation patterns is possible but not considered. In the main we will assume survival times are continuous, or at least the data contain only a small number of ties. Adjustment for lots of ties or the analysis of discrete data is not described.

Throughout we illustrate the methods using one example data set.

Lung cancer data. Results on 272 subjects with 17% censoring. Survival from diagnosis of non-small-cell lung cancer. Six covariates

`Age`: in years

`Sex`: 0=M, 1=F

`Activity score`: 0–4

`Anorexia`: 0=absent, 1=present

`Hoarseness`: 0=absent, 1=present

`Metastases:` 0=absent, 1=present

together with

`Pred:` a subjective prediction of survival time, made by the consulting physician at diagnosis.

2 Basic methods for medical survival analysis

2.1 Survival and hazard functions, likelihood

All of the following should be familiar. Of course we are interested in time $t > 0$ only.

Probability distribution and density functions

$$F(t) = P(T \leq t) \qquad f(t) = \frac{\partial F(t)}{\partial t}$$

Survival function

$$S(t) = P(T > t) = 1 - F(t)$$

Hazard function

$$\alpha(t) = \frac{f(t)}{S(t)} = \text{instantaneous failure rate}$$

Figure 2 has examples of survival distributions and Figure 3 shows the associated survival functions. Which type of plot is most informative?

Cumulative hazard function

$$A(t) = \int_0^t \alpha(u)du$$

Note: $S(t) = \exp\{-A(t)\}$

Likelihood

Under independent right censoring, assumed to have survivor function $G(.)$ and density $g(.)$, we have

$$\text{Likelihood} = \prod_i f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} g(t_i)^{1-\delta_i} G(t_i)^{\delta_i}$$

Usually the censoring mechanism is of no real interest and we only need consider

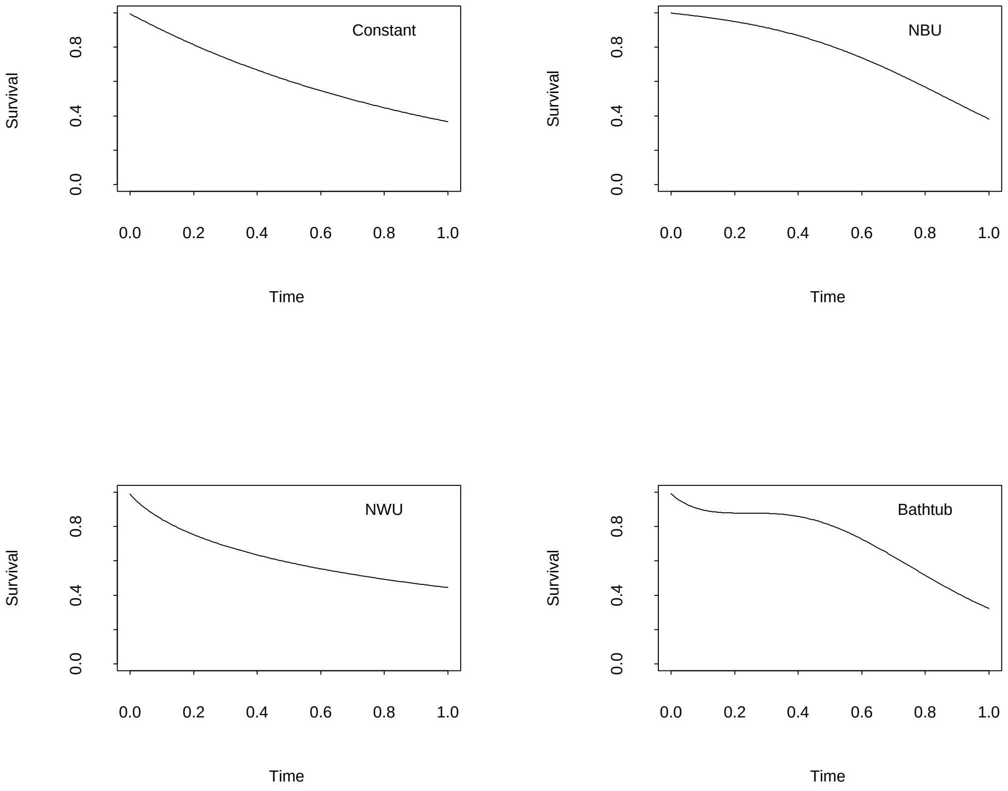


Figure 2: Survivor functions for Figure 3 examples

$$L = \prod_i f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

Strictly this is a *partial likelihood* (not to be confused with another partial likelihood used in fitting proportional hazards).

2.2 Estimating and comparing survival curves

Suppose observation times (in months) are $1, 3, 7^*, 9, 9, 10^*$ where the $*$ indicates a censored time. What is the probability of surviving t months?

Kaplan-Meier estimator

Assume $n(t)$ subjects are *at risk* at time t and $d(t)$ of these fail. Then the estimator is a step function

$$\hat{S}(t) = \begin{cases} 1 & t < \text{smallest observed failure time} \\ \prod_{i:t_i \le t} \left(1 - \frac{d(t_i)}{n(t_i)}\right) & \text{otherwise} \end{cases}$$

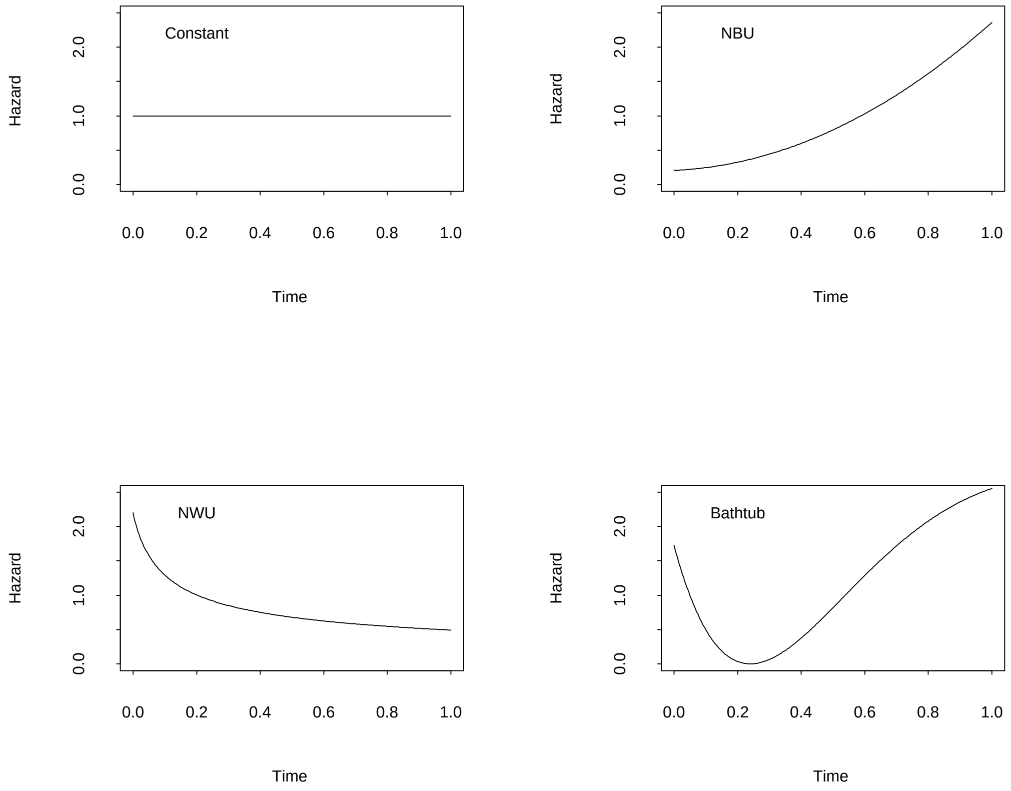


Figure 3: Examples of hazard functions

This is the nonparametric maximum likelihood estimator of survival, sometimes called the product-limit estimator. For the example data

t	$d(t)$	$n(t)$	$1-d(t)/n(t)$	$\hat{S}(t)$
0				
1				
3				
7				
9				
10				

Variance of $\hat{S}(t)$

Greenwood's formula might be used:

$$\hat{\mathrm{Var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{i:t_i \le t} \left(\frac{d(t_i)}{n(t_i)\{n(t_i)-d(t_i)\}} \right)$$

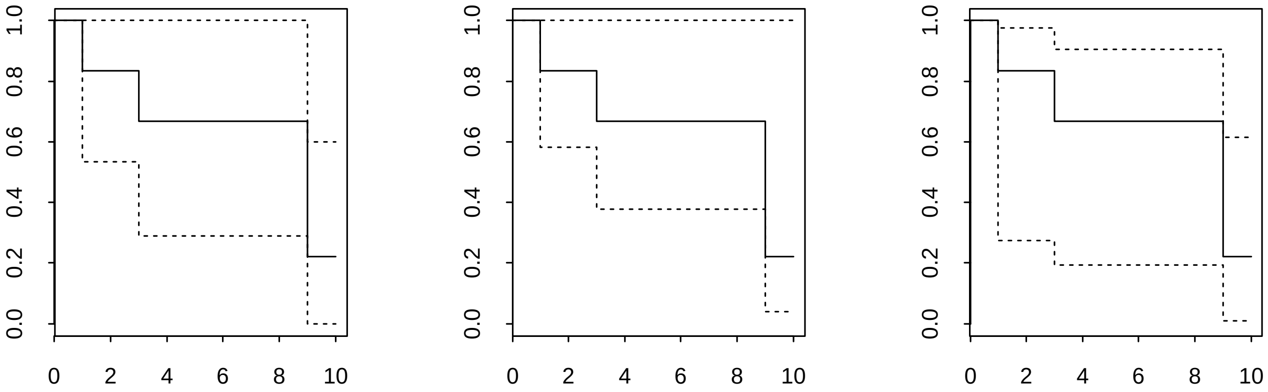
leading to a symmetric CI of the form $\hat{S} \pm 2 \times SE$. In small samples a better CI can be constructed by finding first one for a transformation of $\hat{S}$ and then back-transforming. One possibility is to use $\log(\hat{S}(t))$ and

$$\hat{\text{Var}}\left(\log(\hat{S}(t)\right) = \sum_{i:t_i \le t} \left(\frac{d(t_i)}{n(t_i)\{n(t_i) - d(t_i)\}} \right)$$

This ensures that the cannot include negative values. Another choice is based on $\log\{-\log(\hat{S}(t))\}$ and

$$\hat{\text{Var}}\left(\log\{-\log(\hat{S}(t)\}\right) = \frac{1}{\{\log(\hat{S}(t))\}^2} \sum_{i:t_i \le t} \left(\frac{d(t_i)}{n(t_i)\{n(t_i) - d(t_i)\}} \right),$$

having the advantage that values outside [0,1] are impossible. In small samples the three methods can give different results. For the dummy data, in order untransformed, log, log(-log), the CI's around the Kaplan-Meier estimate are:

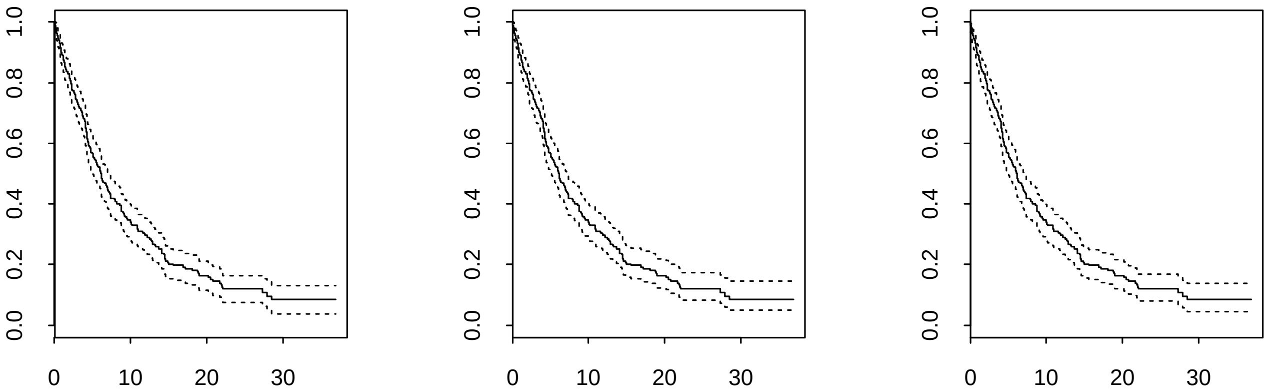


Aside: these were produced by the R commands

```
plot.survfit(survfit(Surv(time,cens),conf.type = "plain"),mark.time = F)
plot.survfit(survfit(Surv(time,cens)), mark.time = F)
plot.survfit(survfit(Surv(time,cens),conf.type = "log-log"),mark.time = F)
```

where the times and censoring data are in `time` and `cens` respectively. Note that R has the sense (unlike many people) not to give estimates outside [0,1], truncating at those values when necessary.

In large samples the choice of method is less important - all lead to almost the same results. Here are the equivalent plots for the lung cancer data:



Comparing survival curves: log rank and related tests

Suppose there are two subgroups of data (perhaps different treatments) and we wish to test a null hypothesis that they have common survival function. A useful class of tests is based on comparisions between observed and expected numbers of failures in each group at each failure time.

Let $n_1(t)$ be the number of subjects in group 1 who are at risk at time t. If $d(t)$ failures are observed at that time the *expected* number from group 1 under the null hypothesis is

$$e_1(t) = d(t) \times \frac{n_1(t)}{n(t)}$$

This is to be compared with the observed number $d_1(t)$ say and a test statistic based on $\{d_1(t) - e_1(t)\}$ is suggested. A variance estimate (don't try to remember this) for $d_1(t)$ is

$$v(t) = \frac{n_1(t)\{n(t) - n_1(t)\}d(t)\{n(t) - d(t)\}}{n(t)^2\{n(t) - 1\}}$$

We now form a weighted sum over failure times to obtain a final test statistic

$$T = \frac{\left[\sum_i w(t_i)\{d_1(t_i) - e_1(t_i)\}\right]^2}{\sum_i w^2(t_i)v(t_i)}$$

which should be χ^2_1 under the hypothesis of equal survival curves. Taking weights $w(t) = 1$ gives the *log rank* test. Other weights give more weight to earlier failure times, including $w(t) = n(t)$ (generalised Wilcoxon) and $w(t) = \hat{S}(t)^\rho$ (the R option, which is Peto's test for $\rho = 1$ and logrank for $\rho = 0$).

Figure 4 shows Kaplan-Meier estimates for the male and female lung cancer data. Using R to test for differences gave the results following and the conclusion of higher female survival for both weighting schemes used.

```
Call:
```

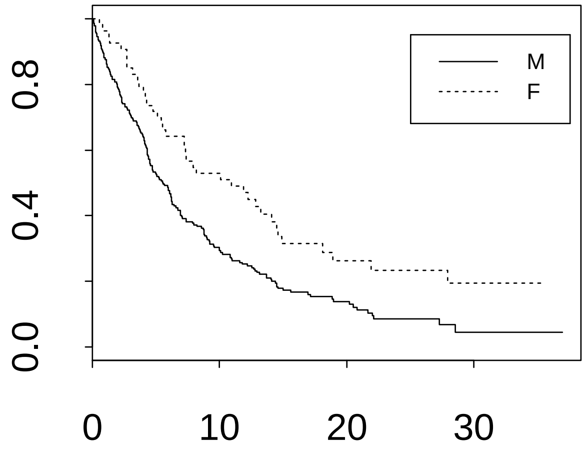


Figure 4: Male and female lung cancer survival

```
survdiff(formula = Surv(lung$time, lung$cens) ~ lung$sex)

             N Observed Expected (O-E)^2/E (O-E)^2/V
lung$sex=0 218      186    165.3      2.59        10
lung$sex=1  54       39     59.7      7.17        10

 Chisq= 10  on 1 degrees of freedom, p= 0.00156


Call:
survdiff(formula = Surv(lung$time, lung$cens) ~ lung$sex, rho = 1)

             N Observed Expected (O-E)^2/E (O-E)^2/V
lung$sex=0 218    110.7     98.5      1.49      9.37
lung$sex=1  54     19.4     31.5      4.67      9.37

 Chisq= 9.4  on 1 degrees of freedom, p= 0.0022
```

These methods extend easily to test for equality of more than two treatment groups.

2.3 Proportional hazards analyses

An overwhelming majority of medical applications of survival analysis involve an assumption (usually untested) of a proportional hazards model.

Cox proportional hazards

$$\alpha(t|x) = e^{\beta x}\alpha_0(t) \qquad S(t|x) = \exp\{-e^{\beta x}A_0(t)\} = S_0(t)^{\exp(\beta x)}$$

where $\alpha_0(t)$ (and therefore $S_0(t)$ and $A_0(t)$) is distribution free.

Assumptions:

- the effects of covariates are multiplicative on the hazard.
- The ratio of hazard functions of two individuals i and j with covariates x_i and x_j, is constant, ie the hazards are proportional.

Partial likelihood estimation

The regression coefficients β can be estimated without the need for any assumptions on the baseline hazard $\lambda_0(t)$ (or equivalently $S_0(t)$) by maximising the *partial likelihood* formed by considering the failure *order* rather than the actual times. Familiarity with this is assumed.

In the absence of ties the partial likelihood is

$$L = \prod_i \left(\frac{e^{\beta x_i}}{\sum_{j\in R_i} e^{\beta x_j}}\right)^{\delta_i}$$

where $R_i = R(t_i)$ is the *risk set* of all subjects still known to be *at risk* at t_i. Adjustment to incorporate ties is possible, though not given here.

The log likelihood, score ($p \times 1$) and information ($p \times p$) are, respectively,

$$l(\beta) = \sum_i \delta_i \left\{\beta x_i - \log(\sum_{j\in R_i} e^{\beta x_j})\right\}$$

$$u(\beta) = \frac{\partial l}{\partial \beta} = \sum_i \delta_i \left\{x_i - \frac{\sum_{j\in R_i} x_j e^{\beta x_j}}{\sum_{j\in R_i} e^{\beta x_j}}\right\} \tag{1}$$

$$i(\beta) = -\frac{\partial^2 l}{\partial \beta^2} = \sum_i \delta_i \left\{\frac{\sum_{j\in R_i} x_j x_j^T e^{\beta x_j}}{\sum_{j\in R_i} e^{\beta x_j}} - \frac{\left(\sum_{j\in R_i} x_j e^{\beta x_j}\right)\left(\sum_{j\in R_i} x_j e^{\beta x_j}\right)^T}{\left(\sum_{j\in R_i} e^{\beta x_j}\right)^2}\right\} \tag{2}$$

Standard likelihood asymptotics apply to the partial maximum partial likelihood estimator $\hat{\beta}$. In particular a test of $H_0 : \beta = \beta_0$ can be performed in three ways:

- Wald test: $(\hat{\beta} - \beta_0)^T \times i(\hat{\beta}) \times (\hat{\beta} - \beta_0)$
- Score test: $u(\beta_0)^T \times i^{-1}(\hat{\beta}) \times u(\beta_0)$
- Likelihood ratio test: $2\{l(\hat{\beta}) - l(\beta_0)\}$

Each of these should be compared with a χ^2 distribution with df equal to the number of coefficients being tested. All three are asympotototically equivalent but there could be differences in small samples. The likelihood ratio test is generally considered the most reliable.

Results of fitting a proportional hazards model to the lung cancer data (from the R function `coxph`) are below. If required, the observed variance matrix $i^{-1}(\hat{\beta})$ can be obtained as the component `$var` of a `coxph` object.

```
              coef exp(coef) se(coef)     z       p
       age  0.0102     1.010  0.00825  1.24 2.2e-01
       sex -0.6850     0.504  0.17968 -3.81 1.4e-04
  activity  0.3457     1.413  0.08220  4.21 2.6e-05
  anorexia  0.3143     1.369  0.14547  2.16 3.1e-02
hoarseness  0.6802     1.974  0.21009  3.24 1.2e-03
metastases  0.4165     1.517  0.22046  1.89 5.9e-02


Rsquare= 0.211   (max possible= 1 )
Likelihood ratio test= 64.4  on 6 df,   p=5.67e-12
Wald test            = 67.2  on 6 df,   p=1.56e-12
Score (logrank) test = 68.7  on 6 df,   p=7.58e-13
```

Note. The Cox model can be used if covariates vary with time, say $x(t)$. All we do is use the appropriate values in the partial likelihood. For example, suppose there are just three subjects and no censoring, but treatment sometimes changed:

Subject	Event time (months)	Treatment
1	6	A always
2	18	A for 1 year, then B
3	30	B for 2 years, then A

Let $x(t) = 0/1$ if A/B, so the partial likelihood is

$$\frac{e^{\beta x_1(t_1)}}{e^{\beta x_1(t_1)} + e^{\beta x_2(t_1)} + e^{\beta x_3(t_1)}} \times \frac{e^{\beta x_2(t_2)}}{e^{\beta x_2(t_2)} + e^{\beta x_3(t_2)}} \times \frac{e^{\beta x_3(t_3)}}{e^{\beta x_3(t_3)}}$$

$$= \frac{e^{\beta \times 0}}{e^{\beta \times 0} + e^{\beta \times 0} + e^{\beta \times 1}} \times \frac{e^{\beta \times 1}}{e^{\beta \times 1} + e^{\beta \times 1}} \times \frac{e^{\beta \times 0}}{e^{\beta \times 0}}$$

Provided there are not too many treatment changes, we represent subjects by *sets* of observations

Subject	New ID	Start	Stop	Event	Treatment
1	1	0	6	1	A
2	2	0	12	0	A
2	3	12	18	1	B
3	4	0	24	0	B
3	5	24	30	1	A

We then include in the partial likelihood only subjects *at risk* at event times

$$= \frac{e^{\beta x_1}}{e^{\beta x_1} + e^{\beta x_2} + e^{\beta x_4}} \times \frac{e^{\beta x_3}}{e^{\beta x_3} + e^{\beta x_4}} \times \frac{e^{\beta x_5}}{e^{\beta x_5}}$$

which is achieved in R via

```
coxph(Surv(start,stop,event)~x)
```

Nelson-Aalen-Breslow estimator

Having obtained $\hat{\beta}$ then the nonparametric maximum likelihood estimator of the cumulative baseline hazard is

$$\hat{A}_0(t) = \sum_{i:t_i \leq t} \frac{d(t_i)}{\sum_{j \in R_i} e^{\hat{\beta} x_j}}$$

This is a step function with increases at observed failure times only. Thus the estimated hazard is zero between failure times and a plot of hazard against time has a very unusual appearance. Smoothing methods can be used if required, though in practice usually it is sufficient to work with the cumulative hazard.

Note that if there are no covariates

$$\hat{A}_0(t) = \sum_{i:t_i \leq t} \frac{d(t_i)}{n(t_i)}$$

The latter is usually called the Nelson-Aalen estimator, the former sometimes called the Breslow estimator.

Having obtained the cumulative hazard the baseline survival function is estimated by

$$\hat{S}_0(t) = \exp\{-\hat{A}_0(t)\}$$

and from this the fitted survival curve for a subject with covariates x_0 can be obtained:

$$\hat{S}(t|x_0) = \hat{S}_0(t)^{\exp(\hat{\beta} x_0)}$$

Again this will be a step function, this time decreasing, and again smooth versions can be obtained. Figure 5 illustrates, showing estimated survival for each activity value at median values of the other covariates (male age 67, no anorexia, hoarseness or metastases).

3 Alternatives to proportional hazards

3.1 Parametric models

There are a large number of fully parametric statistical models available. In medicine most are rarely used, the proportional hazards model so dominates, and so we give (with little comment) details of only a few. In all we only need consider $t > 0$.

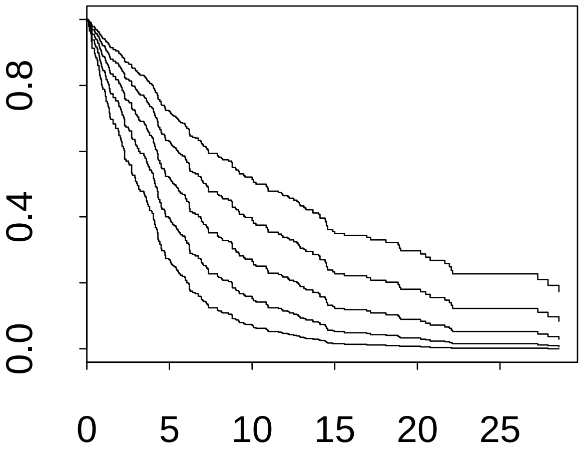


Figure 5: Estimated lung cancer survival by activity score, from 0 (top) to 4 (bottom)

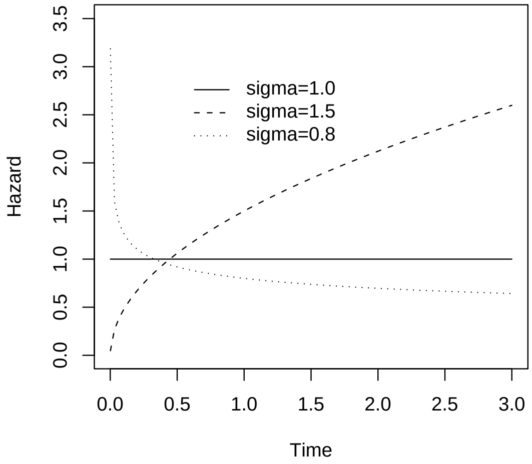
Exponential

$$S(t) = e^{-\alpha t} \qquad \alpha(t) = \lambda \qquad (\lambda > 0)$$

Constant hazard, used mainly in reliability applications. Covariates can be introduced by setting $\alpha = e^{\beta x}$ and including an intercept in x.

Weibull

$$S(t) = e^{-\lambda t^{\sigma}} \qquad \alpha(t) = \sigma \lambda t^{\sigma - 1} \qquad (\lambda > 0, \;\; \sigma > 0)$$



After the semiparametric proportional hazards model this is probably the most widely used. The *shape* parameter σ determines whether the hazard is increasing ($\sigma > 1$), constant ($\sigma = 1$) or decreasing ($\sigma < 1$), as shown in the figure. Covariates can be introduced into the *scale* parameter λ as for the exponential.

Can be parameterised in other ways, such as

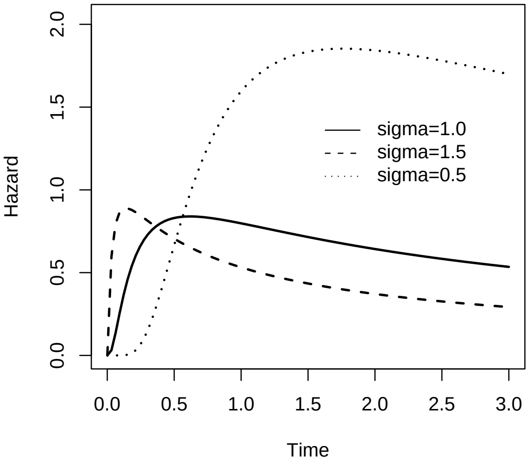
$$S(t) = e^{-(\lambda^* t)^\sigma}$$

Lognormal

Assume that the log of T has a Normal distribution with mean μ and variance σ^2. Then

$$S(t) = \Phi\left(\frac{\log t - \mu}{\sigma}\right)$$

where Φ is the N(0,1) distribution function. Covariates are usually introduced through $\mu = \mu_0 + \beta x$. The figure illustrates the range of hazards this model can describe:

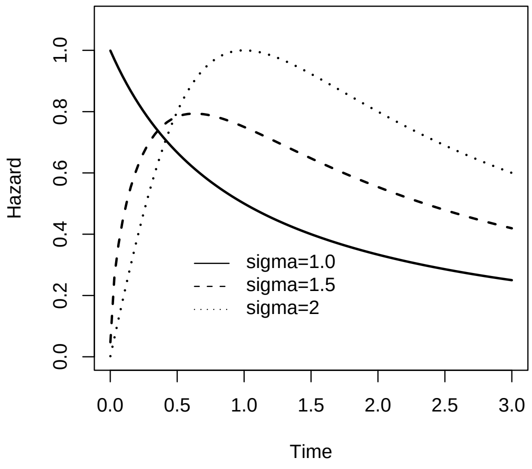


Log-logistic

Here

$$S(t) = \frac{1}{1 + \lambda t^\sigma} \qquad (\lambda > 0, \;\; \sigma > 0)$$

and illustrative hazards are:



3.2 Other families

There are also broad classes of alternatives to the proportional hazards family.

Accelerated failure models

A generalisation of the log-normal. Assume a linear model for the log-lifetime

$$\log T = \mu_0 + \beta x + \epsilon$$

where x does not include an intercept and ϵ is independent of x but otherwise can have any distribution.

Let the survival function of an individual with covariates equal to zero be $S_0(t)$. Then

$$S(t|x) = S_0(t/e^{\beta x})$$

explaining the name. Usually we need to make parametric assumptions about $S_0(t)$ in order to fit this class of model.

Linear models

Under this model there is an assumption that the hazard at time t is a linear combination of covariates, with coefficients which are allowed to vary over time

$$\alpha(t|x) = \alpha_0(t) + \sum_{j=1}^{p} \alpha_j(t) x_j$$

Nonparametric least squares estimation for this class of model is possible, using a method closely related to the *Aalen plots* to be introduced later, which means we do not need to make any assumptions about the form of the $\alpha_j(t)$. A drawback is that there is no restriction to ensure that hazard estimates are always positive.

Note: just as for proportional hazards models, the covariates can be allowed to vary over time.

Piecewise constant hazards

These models are based on an assumption that the hazard remains constant within some specified time intervals. First the time axis is split into intervals (usually pre-defined) $0 = a_0 < a_1 < a_2 < \ldots < a_m = \infty$. The hazard in interval $(a_{j-1}, a_j]$ is α_j.

Proportional odds

In this family it is assumed that the *odds ratio* of survival for two subjects is constant in time, depending only on the covariates. With $S_0(t)$ denoting the baseline survival function the model is

$$\frac{S(t|x)}{1 - S(t|x)} = \frac{S_0(t)}{1 - S_0(t)} e^{-\beta x}$$

Cure models

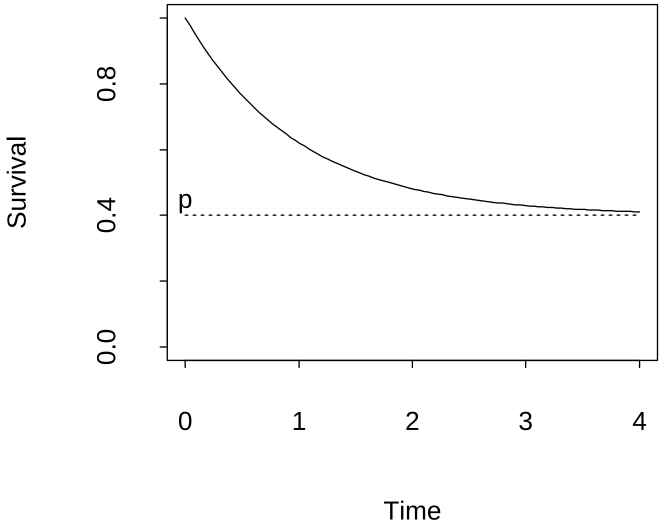


Figure 6: Surviving fractions

An assumption so far has been that $S(t) \to 0$ as $t \to \infty$ so that all distributions are *proper*. The implication is that all subjects will eventually experience the event of interest (even if we don't observe it because of censoring). This is not always a realistic assumption for medical applications in particular - Figure 6 illustrates.

A *mixture* approach assumes that subjects fall into two groups. With probability p there is no chance of the event, and with probability $1 - p$ the event will occur with time-to-event having (proper) survival function $S^*(t)$. Then

$$S(t) = p + (1 - p)S^*(t)$$

Covariates can be introduced by assuming a logistic regression model for group membership, ie

$$P(\text{cure}|x) = p(x) = \frac{e^{\gamma x}}{1 + e^{\gamma x}}$$

and usually a parametric form for $S^*(t)$ is required.

An alternative approach fits an explicitly improper model

$$S(t|x) = \exp\{-\theta F(t)\}$$

where $F(t)$ is a distribution function. The cure fraction is then $p = \exp(-\theta)$ and often we take $\theta = \exp(\beta x)$.

4 Theory based on counting processes

Counting process techniques provide an extremely elegant way of obtaining properties of estimators. In this section we give the main ideas but avoid detail. This means we omit lots of technicalities, including conditions under which the various results hold. In particular we omit conditions on local square integrability (which basically means our processes need finite second moments and things don't go wrong as $t \to \infty$)

4.1 Martingales

Martingale processes

Suppose $M(t)$ is a stochastic process which is continuous to the right. Let $\mathcal{F}_s$ denote the *history* or *filtration* of the process to time s. Then provided

$$E[\mid M(t) \mid] < \infty \qquad \forall\ t \in \mathcal{T} \ \ (\text{range of } t)$$

$M(t)$ is a *martingale* if

$$E[M(t) \mid \mathcal{F}_s] = M(s) \qquad s < t$$

Often we omit the argument t and just talk of the martingale M.

Predictable processes

A process is *predictable* if, just before time t, we know its value *at* time t. A sufficient (but not necessary) condition is left-continuity.

Compensators

$\tilde{X}(t)$ is the *compensator* of the process $X(t)$ if it is predictable and

$$X(t) - \tilde{X}(t)$$

is a martingale, zero at time zero.

Predictable variation

The *predictable variation* process $<M>$ of martingale M is the compensator of M^2, ie

$$M^2 - <M>$$

is a martingale. Roughly,

$$d<M>(t) = \mathrm{var}(\{M(t) - M(t^-)\} \mid \mathcal{F}_{t^-})$$

Core result

If $H(t)$ is predictable then its integral with respect to M,

$$\int_0^t HdM \qquad \left(= \int_0^t H(u)dM(u)\right)$$

is a martingale, with

$$< \int_0^t HdM >= \int_0^t H^2 d<M>$$

Martingale Cental Limit Theorem

Suppose $M^{(n)}$ is a martingale for each of $n = 1, 2, \ldots$ and further that $<M^{(n)}>(t) \to V(t)$, a deterministic function. Then, subject to moment conditions,

$$M^{(n)}(t) \to M^{\infty}(t) \qquad \forall\, t \in \mathcal{T}$$

where M^{∞} is a Normal martingale with $<M^{\infty}>= V$ and $M^{\infty}(t) - M^{\infty}(s) \sim \mathrm{N}(\mathit{0}, V(t) - V(s))$

4.2 Counting processes

The counting process formulation replaces the pair of variables (T_i, δ_i) with the pair of functions $(N_i(t), Y_i(t))$, where

$$N_i(t) = \text{the number of observed events in [0,t] for subject i}$$

$$Y_i(t) = \begin{cases} 1 & \text{subject i is under observation and at risk at time t} \\ 0 & \text{otherwise} \end{cases}$$

Figure 7 shows the N_i and Y_i processes for four hypothetical subjects. The first subject is censored at year 3 and the second has an event at year 4. Subject 3 has multiple events, one at year 0.5 and another at year 2, and is followed to year 4. (The event here is not death, obviously!) Subject 4 is not at risk until year 2 and then experiences an event at year 3.

Note the right-continuity of $N_i(t)$ and the left-continuity of $Y_i(t)$. This disticntion is important. $Y_i(t)$ is a `predictable process`, a process whose value at time t is known infinitesimally before t, at t^- say. $N_i(t)$ is a `counting process`, ie $N_i(t)$ is a stochastic process starting at 0 whose sample paths are right-continuous step functions with jumps of height 1.

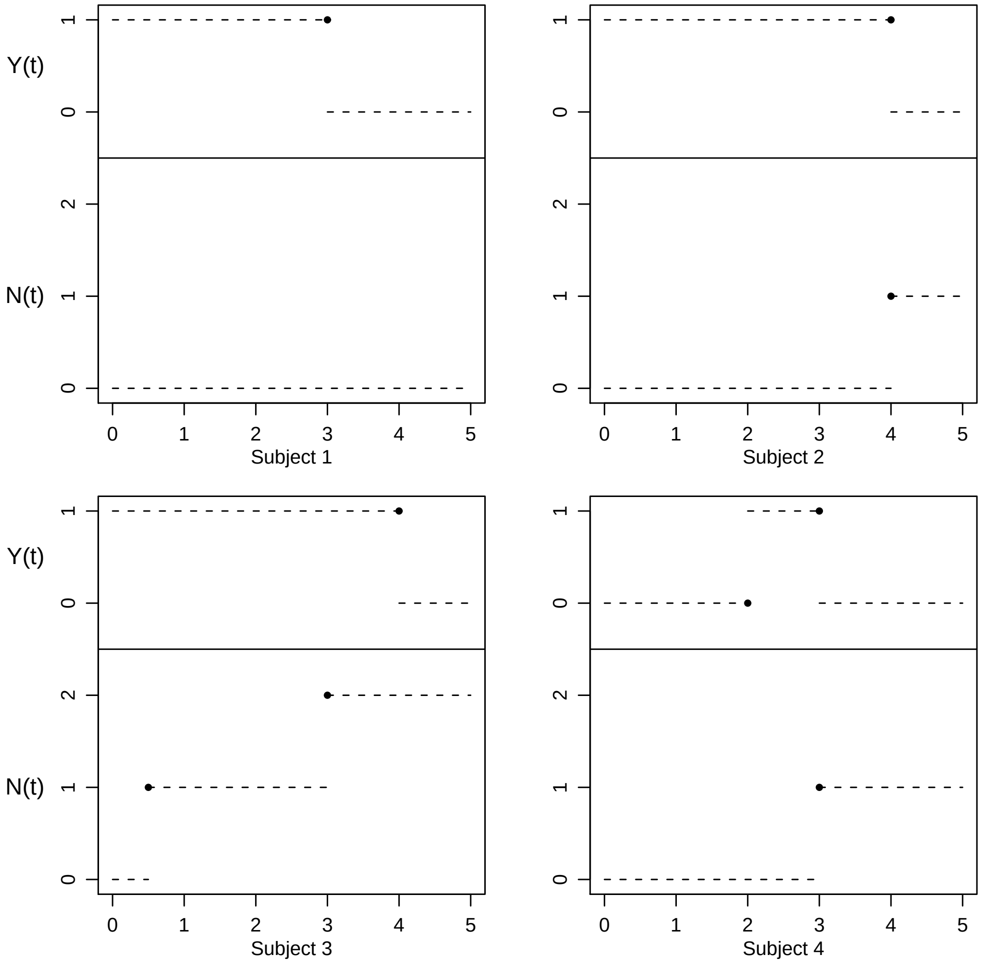


Figure 7: Surviving fractions

For the whole sample we take

$$N(t) = \sum_{i=1}^{n} N_i(t)$$

which is a step function, flat between observed failure times, counting the number of observed events in $[0, t]$, therefore being a counting process.

We also take

$$Y(t) = \sum_{i=1}^{n} Y_i(t)$$

which is the total number of subjects at risk of failure at t.

The counting process formulation includes single event right-censored survival data as a special case

$$N_i(t) = I(T_i \leq t, \delta_i = 1) \text{ and } Y_i(t) = I(T_i \geq t).$$

Note that in this case $N_i(t)$ can take only the values 0 and 1, so the increment in N_i over an infinitesimal time interval

$$dN_i(t) = N_i(t) - N_i(t^-)$$

takes value 1 at failure times, 0 elsewhere (Figure 8 illustrates) so that we can make statements like $T_i = \int_0^\infty t dN_i(t)$ for uncensored data.

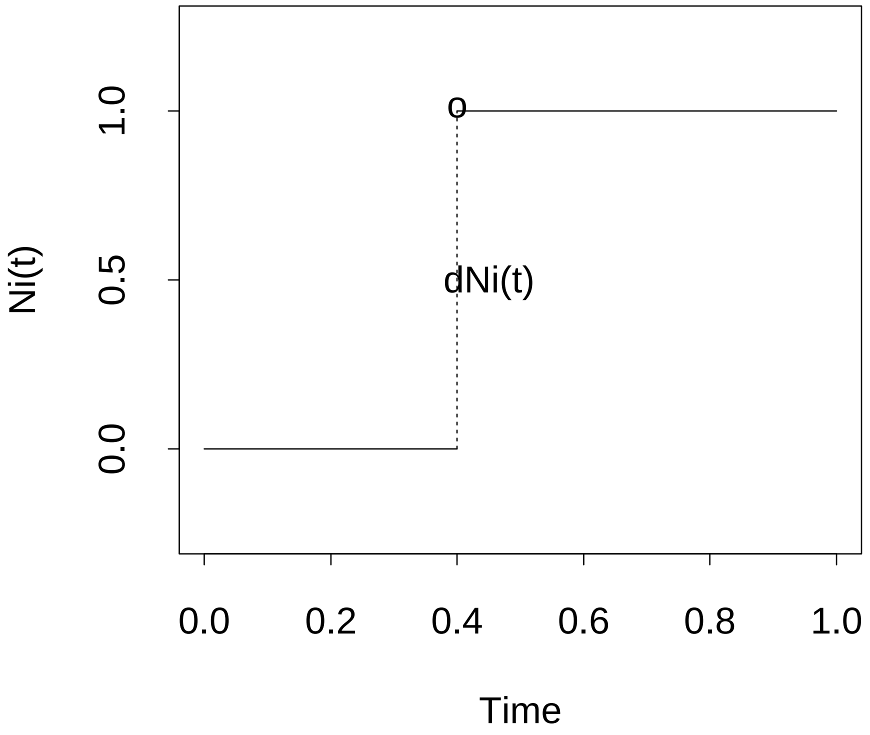


Figure 8: Counting process

This formulation generalizes immediately to multiple events and multiple at-risk intervals, broadening the scope to more elaborate processes. This shifts the emphasis from modelling the hazard of a survival function to modelling the intensity of a point process. We also take

$$Y(t) = \sum_{i=1}^{n} Y_i(t)$$

which is the total number of subjects at risk of failure at t.

Modelling counting processes

In order to develop a statistical model, we need to specify the information on which it is based. For counting process data this is done by specifying the `history`, often called the `filtration`, denoted $\{\mathcal{F}_t; t > 0\}$.

A natural choice is to let $\mathcal{F}_t$ denote the history of the experiment up to and including time t. Generally, $\mathcal{F}_t$ denotes the history of the N_is and any auxiliary processes, such as at risk processes Y_i and, when available, covariate processes X_i. So, for $s \leq t$, $\mathcal{F}_t \in \mathcal{F}_s$, reflecting the increase in information with the passage of time.

To specify the model in terms of this history, note that $\mathcal{F}_{t^-}$ contains all the information on $[0, t)$. Then we have

$$E[dN_i(t) \mid \mathcal{F}_{t^-}] = \lambda(t)dt = Y_i(t)\alpha(t)dt$$

To justify this, note that $dN_i(t)$ can be only 1 or 0, so

$$E[dN_i(t) \mid \mathcal{F}_{t}-] = P(dN_i(t) = 1 \mid \mathcal{F}_{t}-)$$

when subjects are independent, the only component of the history relevant to $dN_i(t)$ is the history of subject i. Moreover, the only relevant part of its history is its status at t^-, $Y_i(t)$. Therefore,

$$P(dN_i(t) = 1 \mid \mathcal{F}_{t}-) = P(dN_i(t) = 1 \mid Y_i(t))$$

. If $Y_i(t) = 0$, then either the event has already happened or the subject is no longer under observation, and $dN_i(t) = 0$ with probability 1. If $Y_i(t) = 1$, the subject is at risk to fail, and

$$P(dN_i(t) = 1 \mid Y_i(t) = 1) = P(t \leq T_i^* < t + dt \mid t \leq T_i^*, t \leq C_i^*),$$

where T_i^* and C_i^* are independent latent failure and censoring times. The independence of T_i^* and C_i^* means that

$$P(t \leq T_i^* < t + dt \mid t \leq T_i^*, t \leq C_i^*) = P(t \leq T_i^* < t + dt \mid t \leq T_i^*) = \alpha(t)dt$$

by the definition of hazard. We can then combine these two possibilities into one equation

$$P(dN_i(t) = 1 \mid Y_i(t)) = Y_i(t)\alpha(t)dt$$

as required.

The interpretation of this is that the expected change in $N_i(t)$ is the product of the event rate, the time period, and the subject's availability to have an event.

The *intensity* of $N(t)$ can be defined extending this argument (roughly) by thinking of $dN(t)$ as a Poisson random variable conditional on the past, with mean

$$E[dN(t) \mid \mathcal{F}_{t}-] = Y(t)\alpha(t)dt = \lambda(t)dt,$$

ie the expected number of events is the product of the event rate, the time period, and the number of subjects available to have an event; and variance

$$\mathrm{Var}(dN(t) \mid \mathcal{F}_{t}-) = \lambda(t)dt$$

The *cumulative intensity* is then

$$\Lambda(t) = \int_0^t \lambda(u)du$$

so that we have $E[N(t)] = \Lambda(t)$.

If covariates are present in a proportional hazards model the intensity is

$$\lambda(t) = \sum_i Y_i(t)e^{\beta x_i}\alpha_0(t)$$

Counting process martingale

Following the above argument, both

$$M(t) = N(t) - \Lambda(t)$$

and

$$M_i(t) = N_i(t) - \Lambda_i(t)$$

can be shown to be martingales with

- $E[dM(t) \mid \mathcal{F}_{t^-}] = 0$
- $E[M(t) \mid \mathcal{F}_s] = M(s)$
- $<M>(t) = \Lambda(t)$

and similar results for $M_i(t)$. The core result above on stochastic integration and the martingale central limit theorem then allow extremely simple (compared to standard methods at least) derivations of properties of most survival estimators. The following two sections provide illustrative examples.

4.3 Example I: the Nelson-Aalen estimator

Recall that, without covariates, the Nelson-Aalen estimator is

$$\hat{A}(t) = \sum_{i:t_i \le t} \frac{d(t_i)}{n(t_i)}$$

Theorem. $\hat{A}(t)$ is asymptotically Normal with mean $A(t)$ and variance which can be approximated by

$$\hat{\text{Var}}(\hat{A}(t)) = \sum_{i:t_i \le t} \frac{d(t_i)}{n(t_i)^2}$$

<u>Proof</u>

In counting-process notation

$$\hat{A}(t) = \int_0^t \frac{1}{Y(u)} dN(u)$$

and on subtracting the cumulative intensity

$$\begin{aligned}
\hat{A}(t) - A(t) &= \int_0^t \frac{1}{Y(u)} dN(u) - \int_0^t \alpha(u) du \\
&= \int_0^t \frac{1}{Y(u)} dN(u) - \int_0^t \frac{1}{Y(u)} Y(u)\alpha(u) du \\
&= \int_0^t \frac{1}{Y(u)} dN(u) - \int_0^t \frac{1}{Y(u)} d\Lambda(u) \\
&= \int_0^t \frac{1}{Y(u)} dM(u)
\end{aligned}$$

So we have $\hat{A}(t) - A(t)$ as the integral of a predictable process with respect to the counting process martingale. The core result immediately applies and we see that $E[\hat{A}(t)] = A(t)$ and

$$<\hat{A}>=\int_0^t \frac{1}{Y(u)^2} d\Lambda(u) = \int_0^t \frac{1}{Y(u)} \alpha(u) du$$

leading to the variance estimator

$$\hat{\text{Var}}(\hat{A}(t)) = \int_0^t \frac{1}{Y(u)^2} dN(u) = \sum_{i:t_i \leq t} \frac{d(t_i)}{Y(t_i)^2} = \sum_{i:t_i \leq t} \frac{d(t_i)}{n(t_i)^2}$$

Moreover the martingale central limit theorem tells us that in large samples we can assume a Normal distribution for $\hat{A}$. (Actually in all of the above we need a restriction to prevent $1/Y(u)$ becoming infinite, but this is not important). □

4.4 Example II: partial likelihood estimation

We now turn to the proportional hazards model. This is a bit more complicated but still a much easier method of obtaining asymptotic properties of $\hat{\beta}$ than any non-counting process technique. The idea is to show that the score function is a martingale and then use a Taylor series to approximate $\hat{\beta} - \beta$ by a multiple of the score. We give an indication of the method, but not full detail.

First, some new notation

$$S^{(0)}(\beta, t) = \sum_{i=1}^n Y_i(t) e^{\beta x_i} \quad S^{(1)}(\beta, t) = \sum_{i=1}^n x_i Y_i(t) e^{\beta x_i} \quad S^{(2)}(\beta, t) = \sum_{i=1}^n x_i x_i^T Y_i(t) e^{\beta x_i}$$

$$E(\beta, t) = S^{(1)}(\beta, t) / S^{(0)}(\beta, t)$$

Theorem. The maximum partial likelihood estimator $\hat{\beta}$ is asymptotically Normal with mean β and variance which can be approximated by the inverse information.

Proof

We begin by considering the score equation (1) as a function of time:

$$U(\beta, t) = \sum_{i:t_i \leq t} \delta_i \left\{ x_i - E(\beta, t) \right\} = \sum_i \left\{ \int_0^t x_i dN_i(u) - \int_0^t E(\beta, u) dN_i(u) \right\}$$

Since

$$\sum_i \int_0^t \left\{ x_i - E(\beta, u) \right\} Y_i(u) e^{\beta x_i} \alpha_0(u) du = 0$$

(check by interchanging the integral and sum) we find

$$U(\beta, t) = \sum_{i=1}^n \int_0^t \{x_i - E(\beta, u)\} dM_i(u)$$

with

$$M_i(t) = N_i(t) - \int_0^t e^{\beta x_i} Y_i(u)\alpha_0(u)du$$

At the true value of β we know M_i is a martingale and the core result applies so that $U(\beta, t)$ is asymptotically Normal with zero mean and predictable variation

$$\begin{aligned} < U(\beta, t) > &= \sum_i \int_0^t \{x_i - E(\beta, u)\}\{x_i - E(\beta, u)\}^T e^{\beta x_i} Y_i(u)\alpha_0(u)du \\ &= \int_0^t \left\{ S^{(2)}(\beta, u) - \frac{S^{(1)}(\beta, u)S^{(1)}(\beta, u)}{S^{(0)}(\beta, u)} \right\} \alpha_0(u)du \end{aligned}$$

We also consider the information equation (2) as a function of time and can show

$$i(\beta, t) = \int_0^t \left\{ S^{(2)}(\beta, u) - \frac{S^{(1)}(\beta, u)S^{(1)}(\beta, u)}{S^{(0)}(\beta, u)} \right\} \frac{1}{S^{(0)}(\beta, u)} dN(u)$$

Recalling the Nelson-Aalen-Breslow estimator, written as $\hat{\alpha}_0(u)du = dN(u)/S^{(0)}(\beta, u)$ we see that asymptotically the information $i(\beta, t)$ will converge to the predictable variation process $< U(\beta, t) >$. With appropriate scaling the martingale central limit theorem also applies and we have an approximation

$$U(\beta, t) \sim \mathrm{N}\left(0, i(\beta, t)\right)$$

Now we remove the limit t and take a Taylor series expansion

$$U(\hat{\beta}) \simeq U(\beta) + i(\beta)(\hat{\beta} - \beta)$$

But $U(\hat{\beta}) = 0$ and so $(\hat{\beta} - \beta) \simeq -i(\beta)^{-1}U(\beta)$ from which, approximately,

$$\hat{\beta} \sim \mathrm{N}\left(\beta, i(\beta)^{-1}\right)$$

Hence we have shown that the maximum partial likelihood estimator has the usual asymptotic properties. □.

5 Diagnostic methods for proportional hazards

Model building is an iterative process. We should always check assumptions and refine the model as necessary. Many tests and graphical procedures are available for the proportional hazards model. Here we provide a small selection only.

5.1 Exploratory plots

Cumulative hazards by strata

In order to check whether a covariate x_0 has a proportional effect on the hazards we can re-fit a *stratified model* with a different baseline for each of the possible values of x_0. (If x_0 is continuous we should group into a small number of categories.) Writing $x_{(0)}$ for the remaining covariates we fit

$$\alpha(t|x_{(0)}, \text{category} j) = \alpha_{0j}(t)e^{\beta x_{(0)}}$$

and obtain the cumulative hazard estimator $\hat{A}_{0j}$ for each category. Plots of $\log\{\hat{A}_{0j}\}$ against time should be roughly parallel if the proportionality assumption holds. Figure 9 illustrates. Comments?

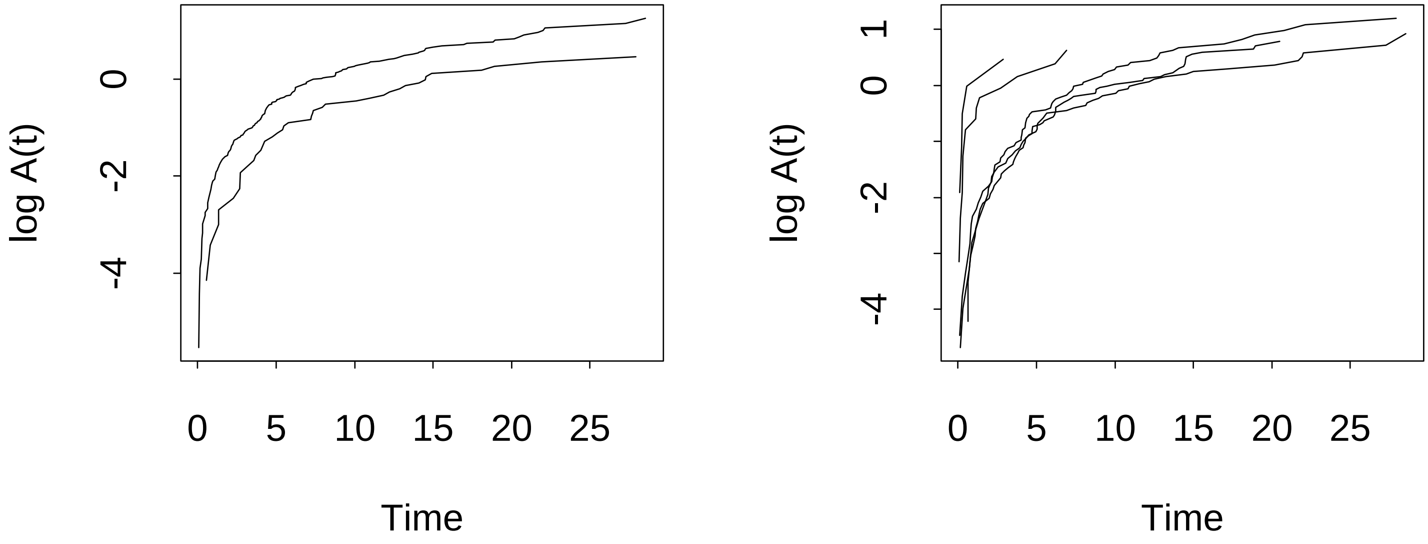


Figure 9: Stratified cumulative hazard plots: sex (left) and activity score (right)

Survival by prognostic index

The *prognostic index* of a subject with covariates x is the linear combination $\hat{\beta}x$. To assess overall goodness of fit we can compare observed survival in groups of subjects with similar prognostic indices as follows.

1. Put the subjects into a small number of groups by prognostic index, perhaps three categories: high, medium and low risk. Try to have reasonably similar group sizes.
2. For each group calculate the mean of the fitted survival curves under the proportional hazards model.

3. Compare with Kaplan-Meier curves for each group.

Figure 10 illustrates.

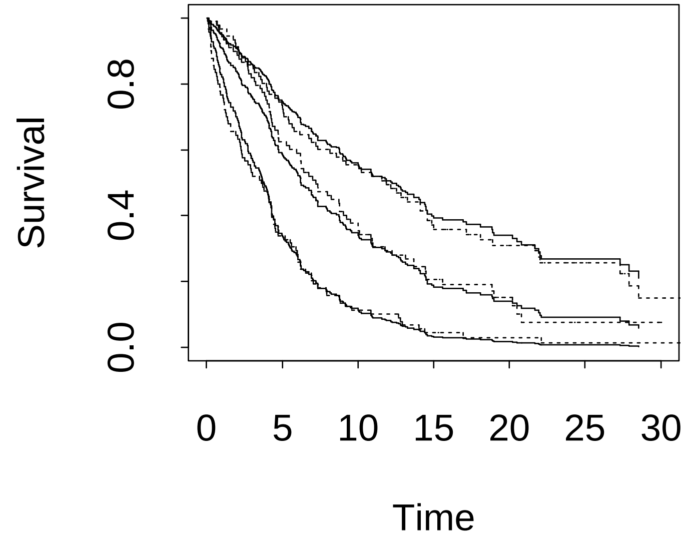


Figure 10: Observed and fitted survival by prognostic index group

5.2 Residuals and influence

In linear regression we are used to checking model adequacy by inspection of plots of residuals between observed and expected responses. In survival analysis residuals defined like this are of little use

- because of the problem of censoring, and
- as the residual distribution will not be Normal, can be highly skew, and will have properties depending upon covariates, in particular different variances.

Several alternative types of residual have been proposed, two of which follow.

Schoenfeld residuals

These are based on the contribution to the score (1), with one vector of residuals at each observed failure time. If case i is observed to fail the corresponding residual is

$$r_i = x_i - \frac{\sum_{j \in R_i} x_j e^{\hat{\beta} x_j}}{\sum_{j \in R_i} e^{\hat{\beta} x_j}} \tag{3}$$

Note that this is a vector, with one component for each covariate. Conditional upon one failure in the risk set R_i the expected value of the covariate of the failure is the right-hand term (exercise), and so the interpretation as a residual is proper. Also $\sum_i r_i = 0$ (why?) as usual for residuals and if the model is appropriate there should be no trends against time. To allow in part for the correlation structure of the residuals a scaled form is often used

$$r_i^* = d \times i(\beta)^{-1} \times r_i$$

where $i(\beta)$ is the information and d is the number of observed failures. The `cox.zph` and `plot.cox.zph` subroutines in R produce and plot these scaled residuals, add smooth estimates of trend to see if there are time effects, and test for correlation with time. The residuals can be plotted against time if required, though usually this leads to a very uneven distribution and a plot against the survival fraction is preferred and is the default in R. (There are also options `transform=''identity''` and `transform=''log''` to plot against time or log(time) is preferred.

Figure 11 illustrates. The associated results of tests for trends are

```
                rho  chisq      p
       age -0.0122 0.0363 0.8489
       sex  0.0358 0.2896 0.5905
  activity -0.0843 1.8916 0.1690
  anorexia -0.1449 4.6720 0.0307
hoarseness  0.0508 0.6034 0.4373
metastases -0.0104 0.0249 0.8746
    GLOBAL      NA 9.7739 0.1345
```

Other residuals

A variety of residuals of one form or another have been suggested for proportional hazards models. These include

- Martingale residuals: differences between *observed* and *expected* numbers of events (given follow-up time) for each subject
- Deviance residuals: a normalised form
- Score residuals: each individual's contribution to the score vector

These are available by `residuals` (or `residuals.coxph)` on a `coxph` object

Case influence

The *case influence* of a subject is a measure of the effect on estimates of deleting him/her from the data set. Letting $\hat\beta_{(i)}$ be the regression estimates when subject i is deleted, a summary is

$$D_i = (\hat\beta_{(i)} - \hat\beta)^T (\text{Var}\hat\beta)^{-1} (\hat\beta_{(i)} - \hat\beta)$$

Usually in reasonable size data sets the deletion of a single subject has no effect on conclusions. However, inspection of D_i can be useful in detecting unusual observations.

Example. Which of the observations below is most influential? All are failures - no censoring.

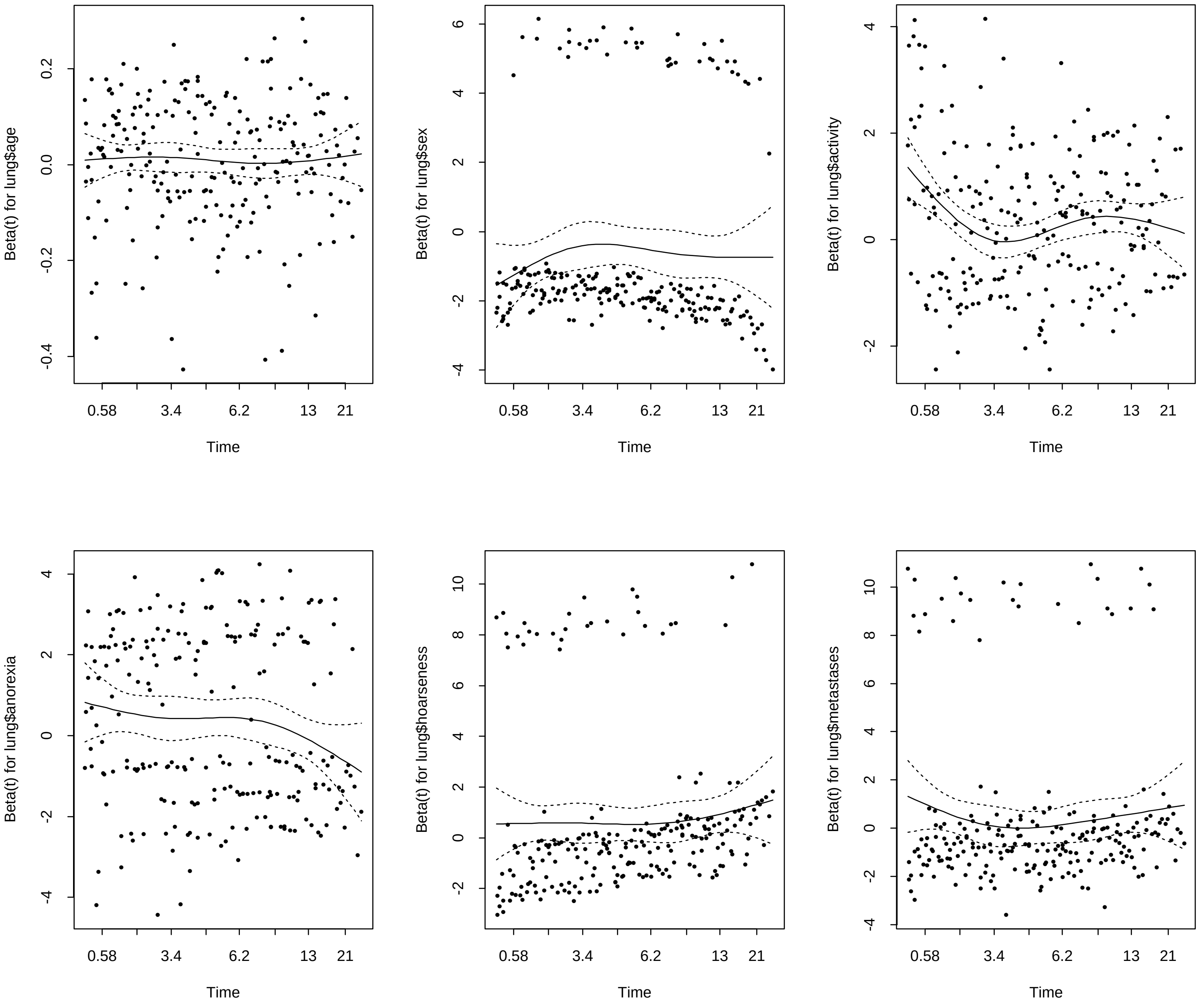


Figure 11: Scaled Schoenfeld residuals

```
i     1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
time 37  42  44  49  52  53  56  61  83  92  93  94 109 131 135 191 400
x     0   0   0   0   0   0   0   0   1   0   1   0   1   0   1   1   1

beta=-1.683 (se=0.672)
```

Figure 12 shows D_i for the lung cancer data. Covariates for the six most influential cases are

```
Subject time   cens age sex activity anorexia hoarseness metastases
    216 14.038    1  65   0        2        1          1          0
    272 36.921    0  76   0        2        1          0          0
    235 16.932    1  57   0        2        1          0          1
    239 17.819    0  53   1        4        0          0          0
    154  6.937    1  73   0        3        1          1          0
    268 30.016    0  59   1        1        1          0          1
```

Comment?

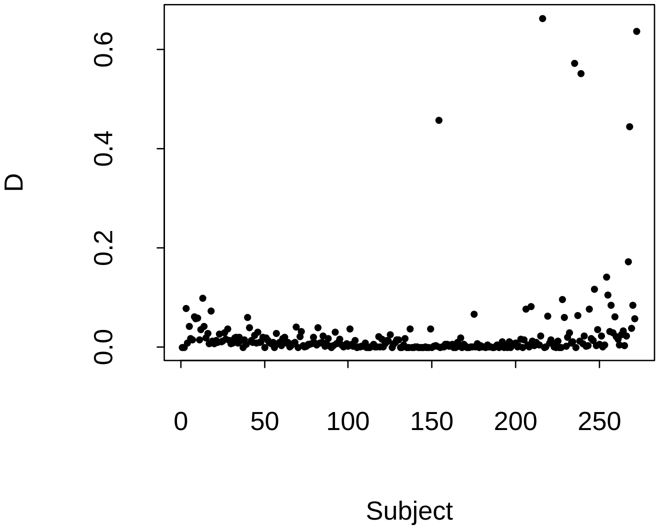


Figure 12: Case influence

5.3 Time varying effects

An assumption so far is that the effect of each covariate is constant in time. Often this is unreasonable - treatment effects wear off, a condition at diagnosis may be important for a while but not later, and so on. Hence we need methods to detect changes in covariate effects and then to fit an adjusted model. One exploratory method is to plot a smoothed version of Schoenfeld residuals as a local estimator $\hat{\beta}$ of the regression coefficient, as seen above. Another useful method is to use Aalen plots.

Aalen plots

These plots are based on the linear hazard model

$$\alpha(t|x) = \alpha_0(t) + \sum_{j=1}^{p} \alpha_j(t) x_j \tag{4}$$

introduced earlier. An estimate of the cumulative regression coefficient

$$A_j(t) = \int_0^t \alpha_j(u) du$$

should be approximately linear if the effect of covariate x_j is constant over time. Even if the true model is proportional hazards the same effect is approximately true, and hence inspection of plots of $A_j(t)$ against t can be used as an exploratory tool to indicate if covariate effects are constant in time.

The basis of the estimation technique is the counting process result that $dN_i(t) \simeq Y_i(t)\alpha(t)$.

1. At each failure time t form a vector dN_t with elements $dN_i(t)$ (ie 1 or 0) for each subject still at risk.
2. Form a design matrix X_t of covariates for these subjects, including an intercept term. Form a vector α_t of regression coefficients $(\alpha_0(t), \alpha_1(t), \ldots)^T$.
3. Equation (4) suggests a linear model $E[dN_t] = X_t \alpha_t$ and hence the estimate
$$\hat{\alpha}_t = (X_t^T X_t)^{-1} X_t^T dN_t$$
provided the inverse exists.
4. Estimate $A(t) = (A_0(t), A_1(t), \ldots)^T$ by
$$\hat{A}(t) = \sum_{u \leq t} \hat{\alpha}_u$$
5. Plot $\hat{A}(t)$ against t and look for changes in slope.

Notes:

- The matrix $X_t^T X_t$ will be singular at the later times t when there may be no variability in a covariate amongst subjects remaining at risk. Usually therefore we stop this procedure some time before the maximum follow-up time.
- The method gives an estimate of the cumulative baseline hazard $A_0(t)$ (under the linear model). This need not be plotted if the purpose (as here) is to look for time varying covariate effects.
- Variance estimates can be obtained but are of little use for this exploratory procedure.
- Figure 13 illustrates for the lung data. There is a suggestion of changes in effect for age, anorexia and perhaps hoarseness.

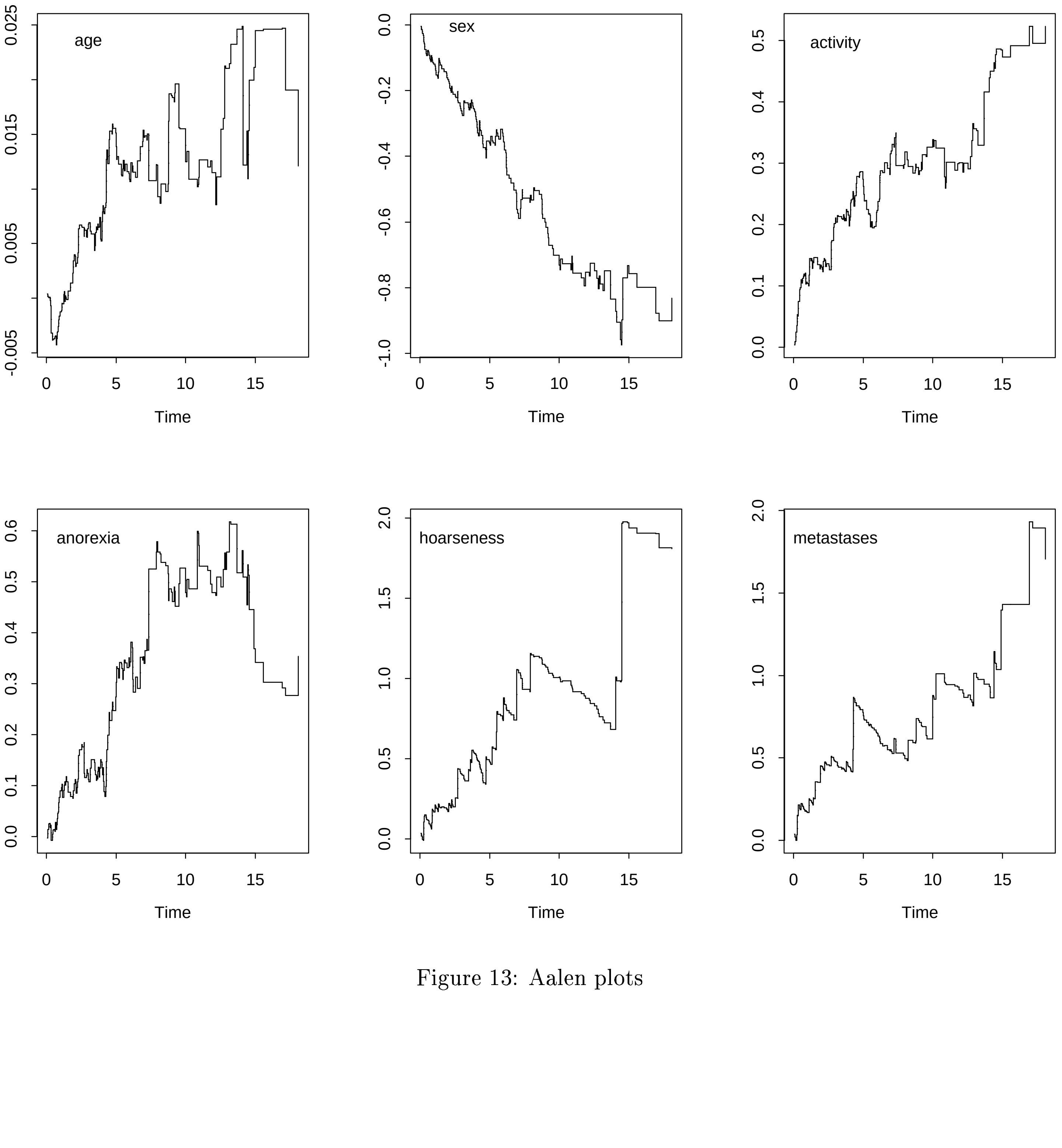


Figure 13: Aalen plots

- If the covariates are highly correlated, it can be dangerous to over-interpret these plots. An apparent change in effect of one covariate may be an artefact due to the effect change of another. It is always best to follow up Aalen plots with a PH analysis allowing either step or smooth changes in regression coefficients.

Allowing time-dependent effects

Having identified a possible change in effect, usually we wish to refit the model with the coefficents β allowed to change at some point τ. This is straightforward.

1. Censor all failure times $> \tau$ and fit a proportional hazards model to give an estimate of β_1, the regression coefficients before τ.
2. Censor all failure times $\leq \tau$ and re-fit to give an estimate of β_2, the regression coefficients after τ.

The combined (log) partial likelihood can be compared with that from a single model to assess whether introducing the changepoint was worthwhile. Several different τ can be attempted and the procedure also works for more than one changepoint (by appropriate censoring).

Example: the lung cancer data with a change at 8 months:

```
Up to 8 months
--------------
                coef exp(coef) se(coef)      z       p
       age  0.00591     1.006   0.0096  0.616 0.54000
       sex -0.65278     0.521   0.2232 -2.925 0.00340
  activity  0.37377     1.453   0.0971  3.848 0.00012
  anorexia  0.55872     1.748   0.1727  3.236 0.00120
hoarseness  0.61215     1.844   0.2276  2.690 0.00710
metastases  0.26255     1.300   0.2639  0.995 0.32000

log lik =-792.65

After 8 months
--------------
               coef exp(coef) se(coef)      z     p
       age  0.0227     1.023   0.0166  1.368 0.170
       sex -0.7717     0.462   0.3064 -2.519 0.012
  activity  0.3536     1.424   0.1613  2.193 0.028
  anorexia -0.3779     0.685   0.3047 -1.240 0.210
hoarseness  0.5529     1.738   0.5986  0.924 0.360
metastases  0.9701     2.638   0.4070  2.384 0.017

log lik =-261.31

Combined log lik = -1053.96  With no change log lik= -1059.103
```

5.4 Prediction and Explained variation

Once a satisfactory fit is obtained, it may be necessary to predict survival for further patients. Obviously the predicted survival curve for a patient with covariates x is $\hat{S}(t|x) = \hat{S}_0(t)^{\exp(\hat{\beta}x)}$. How good are predictions based on this?

One way of assessing is to compare observed and predicted proportions surviving to some fixed time point T^*, a so-called *landmark*, perhaps using cross-validation to avoid using the same data in both fitting and assessing prediction accuracy. For uncensored data we can measure the difference between the *observed survival status* of the individual at T^* and the predicted, $\hat{S}(t|x)$. The observed survival status is 1 if the person has not had the event, 0 otherwise. This means we can measure how good the survival function may be for prediction by forming distances

$$D_i(T^*|x_i) = \begin{cases} 1 - \hat{S}(T^*|x_i) & \text{alive at } T^* \\ \hat{S}(T^*|x_i) & \text{not alive at } T^* \end{cases}$$

and then taking an overall measure

$$BS_X(T^*) = \text{Average}_i \; D_i^2(T^*|x_i),$$

the *Brier score.* (Some authors multiply the above by 2 and some replace the square with $|\;|$). Low values mean a good fit.

Two problems with this idea though are

- choice of T^* is usually arbitrary yet vital
- how to treat information when censoring occurs before T^*.

Methods based on either re-weighting or extrapolation have been suggested to deal with censored data, and an average (or weighted average) over T^* can provide a single summary statistic. However, no one method has yet been generally accepted.

As a consequence, attempts have been made to find a survival equivalent to the linear regression *proportion of variance explained*, R^2, which can be used to compare models and indicate how helpful a model is in predicting future values. $1-R^2$ can be interpreted in many ways, including

- SS(residuals)/SS(Y)
- Var($Y|x$)/Var(Y)
- Squared (multiple) correlation
- E[quadratic loss$|x$]/E[quadratic loss]
- $\{L(0)/L(\beta)\}^{2/n}$ (for Normal data)

All of these have been used as the basis of attempts to define an explained variation measure for survival data under a proportional hazards model, as well as others based on Brier scores. As yet, none have been widely accepted. A likelihood ratio measure

$$R^2_{LR} = 100 \times [1 - \{L(0)/L(\beta)\}^{2/n}] = 100 \times [1 - e^{2\{l(0)-l(\beta)\}/n}]$$

is the simplest and is routinely given in R, but can give an over-optimisitic impression. For the lung cancer data:

Variables	R^2_{LR}
All included	21.1%
Delete age	20.6%
Delete sex	16.1%
Delete activity	16.0%
Delete anorexia	19.7%
Delete hoarseness	18.4%
Delete metastases	20.1%

Notes:

1. Low values (like 21.1%) are typical for proportional hazards models even when the covariates are highly significant.
2. Highly significant covariates should not be taken to mean high explained variation in survival.

6 Frailty

6.1 General

Frailty is the name used in survival analysis for subject-specific unobservable random effects, used to account for

- missing covariates
- measurement error in covariates
- heterogeneity between individuals.

The usual assumption is that a positive-valued random variable Z with pdf $h(.)$ acts multiplicatively on the hazard function. This means (with a proportional hazards model) that the conditional hazard and survival functions are

$$\alpha(t|z,x) = z\alpha_0(t)e^{\beta x} \qquad S(t|z,x) = \exp\{-ze^{\beta x}A_0(t)\}$$

and the marginal (observable) survival distribution is then

$$S(t|x) = \int_0^\infty S(t|z,x)h(z)dz = E_Z[\exp\{-Ze^{\beta x}A_0(t)\}]$$

Notes.

1. Note $\alpha(t|x) \neq \int_0^\infty \alpha(t|z,x)h(z)dz$. Why not?

2. Since Z always appears as a multiple of the hazard there can be identifiability problems. Usually a restriction $E[Z] = 1$ is used to overcome this (provided Z has finite mean - see later).

3. We will assume the censoring mechanism is independent of Z.

4. Ignoring frailty can be shown to lead to underestimation of covariate effects.

6.2 Gamma frailty

This is the most commonly assumed frailty distribution. Let Z be gamma with mean 1 and variance ξ, so that

$$Z \sim \Gamma(1/\xi, 1/\xi) \qquad h(z) = \frac{z^{1/\xi - 1} e^{-z/\xi}}{\xi^{1/\xi}\Gamma(1/\xi)} \qquad (z > 0)$$

Then

$$S(t|x) = \left(\frac{1}{1 + \xi e^{\beta x} A_0(t)}\right)^{1/\xi} \qquad \alpha(t|x) = \frac{e^{\beta x}\alpha_0(t)}{1 + \xi e^{\beta x} A_0(t)}$$

which is a member of the *Burr* family.

Proof

Notes: (don't try to memorise these but make sure you follow the methods).

1. The distribution of frailty amongst both failures at t and survivors at t remains gamma, though with different parameters:

$$[Z|T \geq t] \sim \Gamma(1/\xi, 1/\xi + e^{\beta x} A_0(t)) \qquad [Z|T = t] \sim \Gamma(1/\xi + 1, 1/\xi + e^{\beta x} A_0(t))$$

Proof

2. The marginal survival distribution is more heavily tailed than the conditional survival distribution (ie falls to zero more slowly). Figure 14 illustrates.

3. As time increases the marginal intensity falls relative to the baseline:

$$\frac{\alpha(t|x)}{e^{\beta x}\alpha_0(t)} = \frac{1}{1 + \xi e^{\beta x} A_0(t)} \quad \downarrow \quad \text{as } A_0(t) \uparrow \text{ or } \mathrm{var}(Z) \uparrow$$

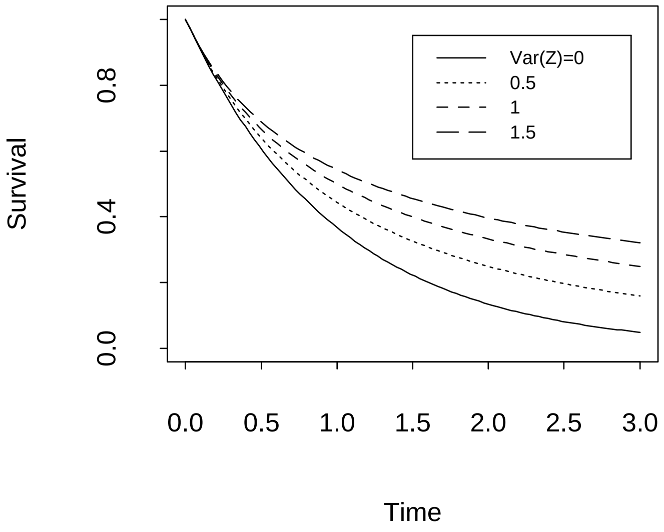


Figure 14: Effect of gamma frailty

4. Marginal hazards converge with time rather than staying proportional. To see this suppose there is only a single binary covariate and let $r = e^{\beta x}$. Let $\alpha_2(t)$ be the hazard in the group with $x = 1$ and $\alpha_1(t)$ be the hazard in the group with $x = 0$. Then in the absence of frailty $\alpha_2(t)/\alpha_1(t) = r$. But if frailty is present

$$\frac{\alpha_2(t)}{\alpha_1(t)} = r\left\{\frac{1+\xi A_0(t)}{1+\xi r A_0(t)}\right\}$$

The rhs is r at $t = 0$ and $\to 1$ as $t \to \infty$.

5. Likelihood estimation with parametric $\alpha_0(t)$ and $A_0(t)$ is straightforward. In the semi-parametric case the partial likelihood method no longer applies (as the baseline hazard does not cancel out). However, if the frailties were observed then we could use partial likelihood for estimation treating them as *offsets*

$$L = \prod_i \left(\frac{z_i e^{\beta x_i}}{\sum_{j\in R_i} z_j e^{\beta x_j}}\right)^{\delta_i}$$

This leads to the use of the *EM algorithm* for estimation with an unspecified baseline, iterating between the following steps.

E-step. Estimate the frailties (and any required functions of them) conditional upon the survival times, using the results at 1 above.

M-step. Estimate the other parameters with the estimates above treated as if they were the observed values.

Further detail is omitted.

6.3 Positive stable frailty

Under gamma frailty the marginal distributions are not now in the proportional hazards family. A frailty distribution which gives proportional hazards marginals is the *positive stable* distribution. This depends on a single parameter ν $(0 < \nu \leq 1)$, has an awfully complex density, infinite moments, and is most easily defined through its Laplace transform

$$E[\exp\{-uZ\}] = \exp\{-u^{\nu}\}$$

Then

$$S(t \mid x) = \exp\{-\{e^{\beta x} A_0(t)\}^{\nu}\} = \exp\{-e^{\nu \beta x} A_0(t)^{\nu}\} = \exp\{-e^{\tilde{\beta} x} \tilde{A}_0(t)\}$$

say. This is of proportional hazards form, but note that now ν and β are not identifiable - we can't distinguish positive stable frailty in a PH model based on β and $A_0(t)$ from no frailty in a PH model based on $\tilde{\beta}$ and $\tilde{A}_0(t)$. Positive stable frailty *is* identifiable in multivariate survival - see next chapter.

6.4 Lognormal frailty

This is the most natural frailty distribution to use if covariates are missing. If x_o are observed covariates and x_m are missing

$$\alpha(t|x) = e^{\beta_m x_m + \beta_o x_o} \alpha_0(t) = Z e^{\beta_o x_o} \alpha_0(t)$$

where $Z = e^{\beta_m x}$. One can argue that there will always be lots of missing covariate information, so the central limit theorem applies and $\beta_m x_m \sim$ Normal, so $Z \sim$ lognormal. The main difficulty with practical use of this frailty distribution is that there is not a closed expression for $S(t \mid x)$.

6.5 Other frailty distributions

A variety of other frailty distributions have been proposed, including: *inverse Gaussian*, the *power variance function* family, and nonparametric *mass point* distributions.

7 Multivariate survival and recurrent events

7.1 Grouped survival data

Sometimes observations form groups or *clusters* with a separate survival time for each cluster member. Often the within-cluster times may not be mutually independent. Examples include

- lifetimes of twins or other siblings
- age first pregnancy of mother and daughter
- time to blindness in each eye under a progressive disease.

For simplicity in this subsection we will assume clusters of size 2 (eg twins) and denote the two survival times by $T_1,\ T_2$. Most of the methods extend to higher dimensions reasonably easily.

Independence working assumption

If the association between survival times is not of interest, we can estimate other parameters by making an *independence working assumption*

$$P(T_1 > t_1,\ T_2 > t_2) = S(t_1, t_2) = S(t_1)S(t_2)$$

This leads to consistent estimation of parameters appearing in $S(t_1)$ and $S(t_2)$, though the usual information-based variance estimates are not appropriate. Instead a robust (or sandwich) estimator of the form

$$A^{-1}BA^{-1}$$

can be used, where A is the usual information (from the second derivative of the log-likelihood or log-partial-likelihood) and B depends on the first derivative - details omitted. This option is available in R by adding `+cluster(subject)` in `coxph`, where `subject` is used to show which survival times are from which person.

Shared frailty models

Positive association is common and can be captured by assuming that subjects within a cluster share a common frailty effect Z, though conditional upon knowing Z their survival times are independent. Under a proportional hazards baseline, and writing $\alpha_j(t) = e^{\beta x_j}\alpha_0(t)$ and $A_j(t) = e^{\beta x_j}A_0(t)$, the joint survival distribution for two cluster members is

$$P(T_1 > t_1, T_2 > t_2) = S(t_1, t_2) = \int_0^\infty S(t_1, t_2|z)h(z)dz = \int_0^\infty S(t_1|z)S(t_2|z)h(z)dz$$

and since $S(t_j|z) = \exp\{-zA_j(t_j)\}$

$$S(t_1, t_2) = E_Z[e^{-Z\{A_1(t_j)+A_2(t_j)\}}]$$

Under $\Gamma(1/\xi, 1/\xi)$ frailty this gives

$$S(t_1, t_2) = \left(\frac{1}{1 + \xi\{A_1(t_j) + A_2(t_j)\}}\right)^{1/\xi}$$

and under positive stable frailty with parameter ν

$$S(t_1, t_2) = \exp\left(-\{A_1(t_j) + A_2(t_j)\}^\nu\right)$$

From these the bivariate densities $f(t_1, t_2)$ can be found by differentiation.

Likelihood and estimation

Let δ_1 and δ_2 denote the censoring/failure indicators for two members of a cluster.

Exercise: write down the likelihood contribution for the cluster.

If a parametric form is assumed for $\alpha_0(t), A_0(t)$ then standard likelihood methods can be used. Under gamma frailty the conditional distribution of Z given the data (ie $t_1, t_2, \delta_1, \delta_2$) is gamma and the EM algorithm can be used to fit a semiparametric model. A *penalised likelihood* method is available in R (or later Splus versions) and is equivalent to EM for gamma frailty. This can also be used for other distributions with closed form densities. With positive stable frailty estimation without assumptions on $\alpha_0(t), A_0(t)$ is more difficult. A number of fairly complex methods have been suggested and research is continuing.

Example. For the lung cancer data we consider the actual and predicted survival times as bivariate responses linked by a shared frailty. To fit this model in R we define

```
  tms:  544 stacked vector of outcomes and predictions

 cens:  same for censoring, all predicted times being considered
        failures

    x:  544x12 matrix for the six covariates, allowed to have different
        effects on outcomes and predictions. Rows 1:272 of form
        (covs, 0) and 273:544 of form (0,covs)

 strt:  272 zeros then 272 ones, used as strata variable to allow
        different baselines for outcome and predictions

pairs:  1:272 then 1:272 to link two repsonses
```

The command

```
coxph(Surv(tms,cens) x+strata(strt)+frailty(pairs))
```

produced (after a little tidying)

```
Outcomes
              coef     se(coef) se2     Chisq  DF  p
age            0.01139 0.0105   0.00850   1.18   1 2.8e-01
sex           -0.78178 0.2382   0.19371  10.77   1 1.0e-03
activity       0.53864 0.1066   0.08812  25.55   1 4.3e-07
anorexia       0.60548 0.1965   0.15436   9.49   1 2.1e-03
hoarseness     0.82256 0.2921   0.22856   7.93   1 4.9e-03
metastases     0.43043 0.3031   0.23979   2.02   1 1.6e-01

Predictions
              coef     se(coef) se2     Chisq  DF  p
age            0.00325 0.0103   0.00824   0.10   1 7.5e-01
sex           -0.26893 0.2308   0.17796   1.36   1 2.4e-01
activity       0.87437 0.1087   0.09051  64.70   1 8.9e-16
anorexia       0.69673 0.1974   0.15357  12.45   1 4.2e-04
hoarseness     0.53636 0.2870   0.22305   3.49   1 6.2e-02
metastases     0.89023 0.2944   0.22983   9.15   1 2.5e-03

frailty(pairs)                          368.95 139 0.0e+00
```

```
Variance of random effect= 0.594   I-likelihood = -2236.9
Degrees of freedom for terms=   8.6 138.9
Likelihood ratio test=593  on 148 df, p=0  n= 544
```

Discuss

Copula models

One difficulty with the shared frailty approach is that the parameters and shape of the frailty distribution simultaneously affect two things:

- the marginal survival distributions $S_1(t_1)$ and $S_2(t_2)$, and
- the association between T_1 and T_2.

A way round this problem is to make separate assumptions about these two features. First, assume the forms of the marginal survival distributions, eg proportional hazards. Second, write the joint distribution as a function (a *copula*) of the marginals. For instance

$$S(t_1, t_2) = S(t_1)S(t_2)$$

ie independence (trivial), or

$$S(t_1, t_2) = \left[S(t_1)^{1-c} + S(t_2)^{1-c} - 1\right]^{1/(1-c)} \qquad (c > 1)$$

which allows a single parameter c to determine the association. As yet copula models have had little use in practice.

7.2 Measuring association

How can association between two survival times T_1 and T_2 be measured? A variety of both *global* and *local* association measures have been proposed.

First, note that often it makes sense to remove marginal effects by first transforming to uniformity. This means that we investigate association between the uniformly transformed variables

$$U_1 = 1 - S_1(T_1) \qquad U_2 = 1 - S_2(T_2)$$

rather than T_1 and T_2 directly. Since each U_i has $U(0,1)$ distribution any covariate effects are removed, and extreme values arising from skew distributions of T_1 and T_2 have reduced influence. Figure 15 illustrates.

Global Measures

These are overall single-number summaries of the association. Examples include

1. Correlation coefficient (as usual)

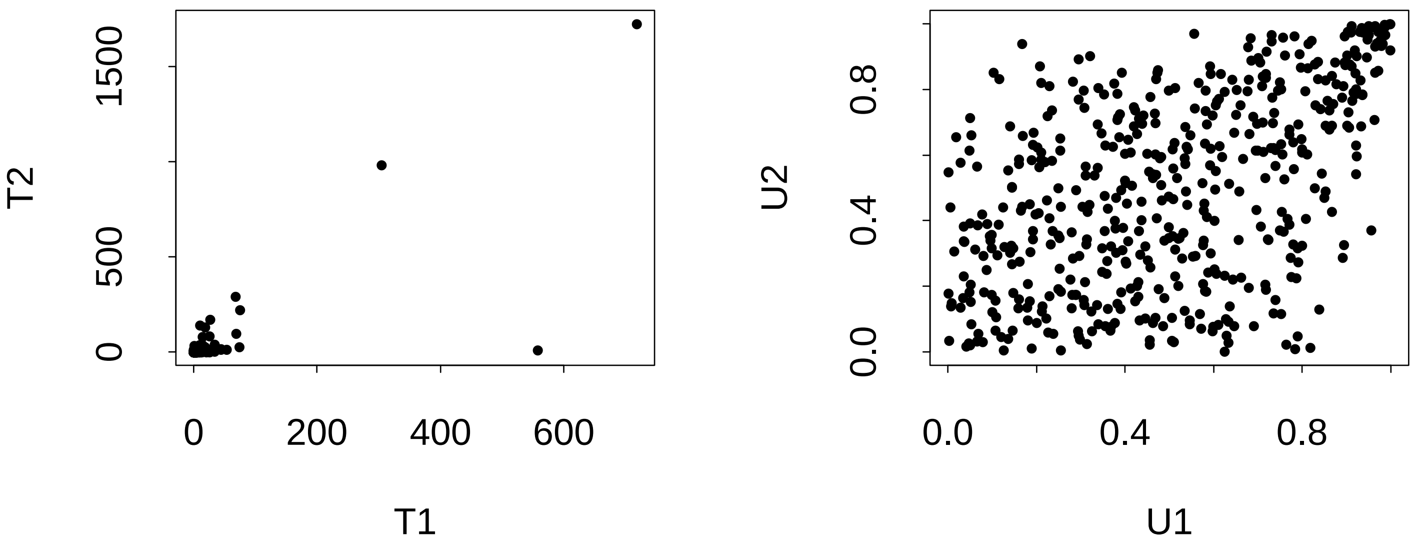


Figure 15: Marginally transforming survival data

2. Kendall's τ:

$$(2P\{(T_{11} - T_{21})(T_{12} - T_{22})\} > 0) - 1$$

3. Spearman's rank correlation (as usual).

For each, estimation in practice is complicated by censoring in the data.

Local measures

Often a single number is too simplistic since association can vary with time. For instance, in Figure 15. knowing that U_1 is low may tell us very little about U_2, whereas knowing U_1 is large tells us U_2 is likely to be large. Local measures take this into account.

Probably the most widely used is Oakes' θ:

$$\theta^*(t_1, t_2) = \frac{\alpha_1(t_1|T_2 = t_2)}{\alpha_1(t_1|T_2 > t_2)} = \frac{S(t_1, t_2) \times \frac{\partial^2 S(t_1,t_2)}{\partial t_1 \partial t_2}}{\frac{\partial S(t_1,t_2)}{\partial t_1} \times \frac{\partial S(t_1,t_2)}{\partial t_2}}$$

Exercise: check the second and third terms are the same.

For shared gamma frailty this reduces for all t_1 and t_2 simply to 1+Var(Z) (check this).

Another measure, with a more intuitive interpretation, and which should be used only after marginal transformation, is Lehmann's c

$$c_1^*(u_1, u_2) = P(U_1 > u_1|U_2 > u_2)/P(U_1 > u_1)$$

or

$$c_2^*(u_1, u_2) = P(U_1 < u_1 | U_2 < u_2)/P(U_1 < u_1)$$

Often this is measure is considered only on the diagonal and is defined by $c(u) = c^*(u, u)$.

Figure 7.2 illustrates for a model in which T_1 and T_2 each have gamma frailties (Z_1 and Z_2) which are *correlated* rather than shared, with $\text{Corr}(Z_1, Z_2) = \rho$. Further details of this model are omitted.

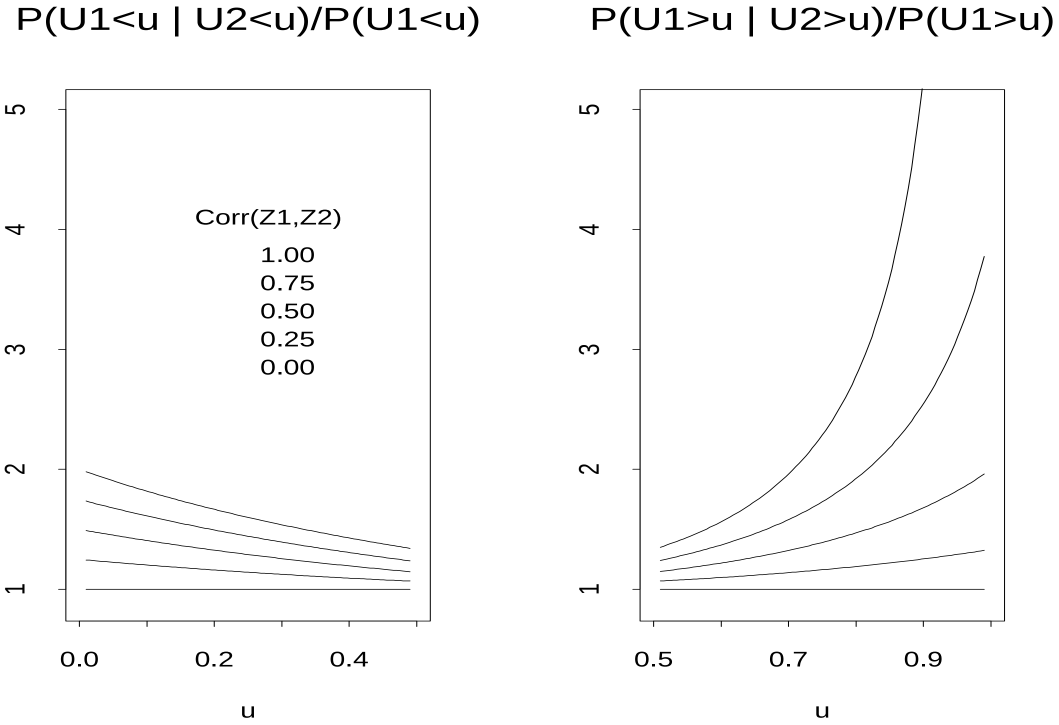


Figure 16: Lehmann's c for correlated frailties

7.3 Recurrent events

A different form of multivariate problem occurs when we have recurrent events, such as epileptic seizures, heart attacks, or infections for dialysis patients. Times between events are now of interest and we have several observations on each subject, usually mutually associated.

Figure 17 illustrates, showing the times at which patients requested morphine painkiller after surgery, for an illustrative sample of 5 patients.

One approach to this type of problem is to analyse the times between events separately. For instance, in tumour recurrence studies we might

- start by analysing only the times to first recurrence
- next analyse the times between first and second recurrence separately, using time to first as a covariate perhaps

and so on. This is useful if there are only few events per person. If not, an alternative is to extend survival methods and think of $\alpha(t)$ as being the *intensity* of points rather than a hazard

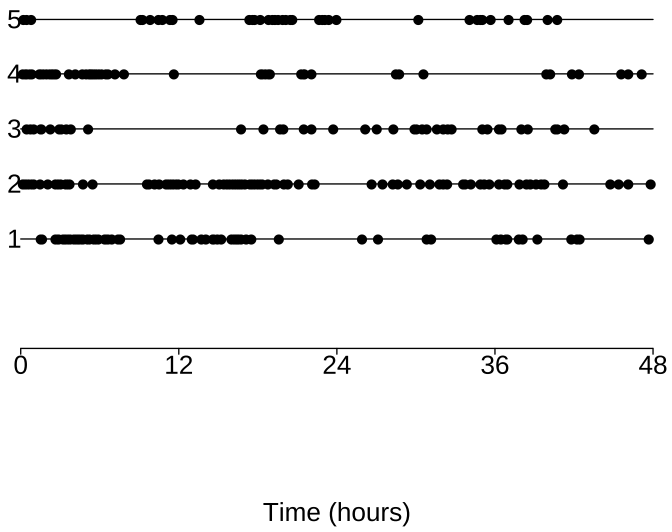


Figure 17: Morphine requests after surgery for sample of five patients

and acknowledge that subjects can have more than one event. The Nelson-Aalen-Breslow estimator of cumulative intensity still works, and counting process theory applies immediately ($N(t)$ is the cumulative number of observed events). The likelihood contribution of a subject with d events, at times $t_1, t_2, \ldots, t_d$, in an observation window $[0, \tau]$ (assuming continuous time), is

$$\left\{\prod_{j=1}^{d} \alpha(t_j)\right\} \exp\{-\int_0^{\tau} \alpha(u) du\}$$

A proportional hazards model to allow covariate effects can be assumed, with estimation under a partial likelihood formed in the usual way - comparing covariates of subjects who experience an event at time t with those at risk at that time.

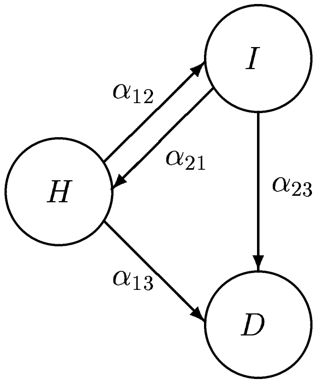
$$\prod_{\text{all event times } t} \left(\frac{\sum_{i\text{:event at } t}(e^{\beta x_i})}{\sum_{j\text{:at risk at } t}(e^{\beta x_j})} \right)$$

The only real difference is that subjects do not fall out of the risk set once the event is observed. The `start, stop` option in `coxph` can be used to structure the data for this type of analysis. Moreover, frailty effects can be included just as for univariate and grouped survival: each subject has their own frailty which acts multiplicatively on the hazard.

If the baseline intensity is assumed to be constant in time then often an analysis will be based only on the *counts* of how many events occur in give periods, the actual event times perhaps not being of interest. If the baseline varies in time then usually we need some natural starting point (such as recovery from surgery) from which everything is referenced. Extensions which allow the intensity to vary with time since last event are also possible.

7.4 Multistate models

A problem which is mathematically similar but conceptually different allows subjects to move between various *states.* For example we might have three states: healthy (H, state 1), illness (I, state 2) and death (D, state 3):



$\alpha_{31} = \alpha_{32} = 0$

Keeping to continuous time, the *transition intensity* $\alpha_{ij}(t)$ gives the instantaneous hazard for movement out of state i and into state j $(i \neq j)$. Usually this is assumed to be Markov, ie depending only upon covariates and the states i and j, rather than history of earlier transitions. The total intensity for leaving state i is $\sum_{j \neq i} \alpha_{ij}(t)$, denoted $\alpha_i(t)$. The likelihood is then easy to build up. For instance, suppose a subject starts at time 0 in state 1, enters state 2 at time t_1, and moves to state 3 at time t_2, remaining there until observation is complete at time τ (this needn't now be healthy/disease/death). The likelihood contribution from this subject is

$$\exp\left\{-\int_0^{t_1} \alpha_1(u)du\right\} \alpha_{12}(t_1) \exp\left\{-\int_{t_1}^{t_2} \alpha_2(u)du\right\} \alpha_{23}(t_2) \exp\left\{-\int_{t_2}^{\tau} \alpha_3(u)du\right\}$$

Notes:

1. If a state is absorbing then of course $\alpha_i(t) = 0$ and there is no further likelihood contribution once the state is entered.

2. Often the intensities are assumed to be constant in time, greatly simplifying the analysis.