Chapter 3

The Cox Proportional Hazards Model

3.1 Overview of the Cox proportional hazards model

3.1.1 Introduction

In the last chapter we considered testing for a difference in survival based on a categorical covariate, such as sex. This lets us know if there is a difference, but it doesn’t help us answer how much more at risk one individual is than another. Similarly, it is not ideal when dealing with a continuous covariate: we can arbitrarily bin the covariate into groups, but a different grouping will give a different test result. Importantly, it does not adjust for confounders, making it difficult to justify for observational studies.

In this chapter, we seek to achieve two things:

- incorporate continuous covariates into our survival analysis; and
- analyse the effect of (not just the presence of) covariates on survival.
We do these within a unified framework, namely using Cox’s proportion hazards model (PHM).

### 3.1.2 Who is Cox?

David Cox is an English statistician, and a reknowned one at that. He has written over 300 papers or books on a variety of topics, has advised government, was knighted for his contribution to science, and holds numerous fellowships and awards. His paper introducing the proportional hazards assumption and inference for it (Cox, 1972), has been cited around 20 000 times, according to Google scholar.

### 3.1.3 Recall: linear regression

In linear regression, we use a predictor variable $x$ to explain some of the uncertainty in a response variable $y$. If the $i$th individual has independent and dependent variables $x_i$ and $y_i$, respectively, the linear model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$. Note that this is a model, and it depends on certain assumptions, e.g. that the relationship is linear and errors are Gaussian. Note also that as $\epsilon_i \in (-\infty, \infty)$ and $\beta_0 + \beta_1 x \in (-\infty, \infty)$, so must $y_i \in (-\infty, \infty)$.

### 3.1.4 Survival regression

How can we do likewise for survival data? We choose to focus on models for the hazard function, as this allows statements such as “the risk to males is $X$ times the risk to females” more readily than using the survival function as our basis.

A natural first guess for such a regression survival model would be

$$h(t, x) = \beta_0 + \beta_1 x.$$
There is no “error” term as the randomness is implicit to the survival process. Here we have used the notation $h(t, x)$ to be the hazard function for an individual whose “independent” variable has the value $x$, while $\beta_0$ is a baseline hazard function (for the time being assumed constant in time $t$) for individuals with $x = 0$.

However, this is a bad model. The range of $\beta_0 + \beta_1 x$ may extend below zero for certain values of $\beta_1$ or $x$, but the range of $h(t, x)$ must be $[0, \infty)$.

Luckily, a similar problem has arisen and been solved in generalised linear modelling. There, the predictors are incorporated into different distributions for the dependent variable. For a Poisson model, the mean must be positive, and the exponential function is used as the canonical link function between covariates and mean. We thus follow suit by exponentiating the covariate terms:

$$h(t, x) = \exp(\beta_0 + \beta_1 x) = h_0 \exp(\beta_1 x) > 0.$$  

We can have more than one predictor if we use matrix notation:

$$h(t, \mathbf{x}) = h_0 \exp(\mathbf{\beta}^T \mathbf{x}).$$

Note that for a cohort with identical predictors $\mathbf{x}$, the above form implies that lifetimes are exponentially distributed, which we know to be unrealistic.

### 3.1.5 The Cox proportional hazards model

We therefore consider the following generalisation:

$$h(t, \mathbf{x}) = h_0(t, \mathbf{\alpha}) \exp(\mathbf{\beta}^T \mathbf{x}),$$

where $\mathbf{\alpha}$ are some parameters influencing the baseline hazard function.

Note that we have decomposed the hazard into a product of two items:

- $h_0(t, \mathbf{\alpha})$, a term that depends on time but not the covariates; and
- $\exp(\mathbf{\beta}^T \mathbf{x})$, a term that depends on the covariates but not time.
This is the Cox PHM. The beauty of this model, as observed by Cox, is that if you use a model of this form, and you are interested in the effects of the covariates on survival, then you do not need to specify the form of $h_0(t, \alpha)$. Even without doing so you may estimate $\beta$. The Cox PHM is thus called a semi-parametric model, as some assumptions are made (on $\exp(\beta^T x)$) but no form is pre-specified for the baseline hazard $h_0(t, \alpha)$.

To see why it is called the PHM, consider two individuals with covariates $x_1$ and $x_2$ (which we can treat for simplicity as scalars). Then the ratio of their hazards at time $t$ is

$$\frac{h(t, x_1)}{h(t, x_2)} = \frac{h_0(t, \alpha)}{h_0(t, \alpha)} \exp(\beta x_1) \frac{h_0(t, \alpha)}{h_0(t, \alpha)} \exp(\beta x_2) = \exp(\beta (x_1 - x_2)).$$

In other words, $h(t, x_1) \propto h(t, x_2)$, i.e. the hazards are proportional to each other and their ratios does not depend on time. In particular, the hazard for the individual with covariate $x_1$ is $\exp\{\beta(x_1 - x_2)\}$ times that of the individual with covariate $x_2$. This term, $\exp\{\beta(x_1 - x_2)\}$, is called the hazard ratio comparing $x_1$ to $x_2$.

If $\beta = 0$ then the hazard ratio for that covariate is equal to $e^0 = 1$, i.e. that covariate doesn’t affect survival. Thus we can use the notion of hazard ratios to assess if covariates influence survival. The hazard ratio also tells us how much more likely one individual is to die than another at any particular point in time. If the hazard ratio comparing men to women were 2, say, it would mean that, at any instant in time, men are twice as likely to die than women.

Note however that this is a model—it could be wrong. There may be an interaction between covariates and time, in which case hazards are not proportional. In the next chapter we learn how to check for violations of the proportional hazards assumption and in the chapter that follows that we extend the PHM to incorporate such interactions. Similarly, there is no reason why we should expect the log of the hazard function to be linear in the covariates. Unlike in linear regression, there is no simple way to absorb additional variation (the $\sigma$ term) as the stochasticity in the data is generated implicitly by the hazard function, so that missing a predictor out cannot really be
rectified by hoping the error term can capture that source of variation. For now, we consider the proportional hazards assumption to be appropriate.

We consider the following special cases one at a time:

- one continuous covariate;
- two continuous covariates;
- one binary covariate;
- one categorical covariate; and
- one continuous and one categorical/binary covariate.

Note, though, that the approach generalises to more complex models in an obvious way. Each will be illustrated by an example of an analysis of survival data using the uis.dat data. These relate to the length of time drug users are able to avoid drug use following a residential treatment programme. Eight covariates were also recorded. Refer to Hosmer et al. (2008) for more details and references. In all these examples, $t_i$ is the survival time (until first reuse of drugs or leaving the study [note that it was considered likely that those leaving the study had taken up drug use again and so these are not considered to be right censored in this particular example]) or the time of an (unspecified) right-censoring event.

### 3.1.6 A single continuous covariate

- The covariate is $x \in \mathbb{R}$.
- The parameter is $\beta \in \mathbb{R}$.
- The hazard rate is $h(t, x) = h_0(t) \exp(\beta x)$.
- The hazard ratio for two individuals with covariates $x_1$ and $x_2$ is $\exp\{\beta(x_1 - x_2)\}$. Increasing $x$ by one unit scales the hazard rate by $\exp\{\beta(x + 1 - x)\} = e^\beta$. We can thus interpret $\beta$ as the increase in log hazard per unit of $x$. 
Example: age of drug addicts. Let $a_i$ be the age of drug addict $i$ at the time of admission to the programme. The fitted hazard rate is

$$h(t, a) = h_0(t) \exp(-0.013a).$$

(To learn how it was fitted, see later.) Thus, each year of an addict's age is estimated to multiply the risk of taking drugs again by $e^{-0.013} = 0.99$. This turns out not to be statistically discernible from zero ($p = 0.07$), and so we have no reason to reject the hypothesis that age has no effect on reversion to drug use (a standard response to a $p$-value approaching 5% would be to recommend more data be collected, but the sample size of 600 is already fairly large).

### 3.1.7 Two continuous covariates

*Using scalar notation:*

- The covariates are $(x_1, x_2) \in \mathbb{R}^2$.
- The parameters are $(\beta_1, \beta_2) \in \mathbb{R}^2$, or $(\beta_1, \beta_2, \beta_{12}) \in \mathbb{R}^3$ if there is an interaction between $x_1$ and $x_2$.
- With no interaction, the hazard rate is $h(t, x_1, x_2) = h_0(t) \exp(\beta_1x_1 + \beta_2x_2)$.
- With an interaction, the hazard rate is $h(t, x_1, x_2) = h_0(t) \exp(\beta_1x_1 + \beta_2x_2 + \beta_{12}x_1x_2)$.
- With no interaction, the hazard ratio for two individuals with covariates $(x_1^1, x_2)$ and $(x_1^2, x_2^2)$ is $\exp\{\beta_1(x_1^1 - x_1^2)\}$. Increasing $x_1$ by one unit while keeping $x_2$ fixed scales the hazard rate by $e^{\beta_1}$. A similar interpretation holds for $\beta_2$ by symmetry.
- With an interaction, the $\beta$s can no longer be interpreted thus. The effect of increasing $x_1$ while keeping $x_2$ fixed depends on the value of $x_2$. 
Example: number of previous drug treatments and depression of drug addicts. Let $b_i$ be the number of previous drug treatments and $e_i$ be the Beck depression score for individual $i$. Here, $b_i$ is in the range $[0, 40]$ and $e_i$ in $[0, 54]$. The fitted hazard rate for the no-interaction model is:

$$h(t, b, e) = h_0(t) \exp(0.030b + 0.010e).$$

A Beck scale of 0–9 indicates no depression, 10–18 indicates mild–moderate depression, 19–29 indicates moderate–severe depression and 30–63 indicates severe depression (Beck et al., 1988). Thus the hazard ratio for a “typical” addict with severe depression (40) relative to one with mild depression (15) is $e^{25 \times 0.01} = 1.27$. The $p$-value for the hypothesis that there is no effect based on depression is just under 5%, indicating some weak evidence of an effect, with more depressed drug users slightly more likely to go back to drug use.

The hazard ratio for someone who has already undergone 5 treatments for drug addiction (the mean for these data) relative to someone who has never had any treatment is $e^{5 \times 0.03} = 1.16$. A serial user with 20 prior treatments has a hazard rate $e^{20 \times 0.03} = 1.8$ times that of someone undergoing his/her first treatment. This is highly significant, with the $p$-value for the hypothesis that there is no effect of prior treatment being less than 0.01%.

The fitted hazard rate for the interaction model is:

$$h(t, b, e) = h_0(t) \exp(0.0079b + 0.0043e + 0.0012be).$$

None of these terms is now significant, although the model as a whole is. We would therefore throw away the interaction term and go back to the no-interaction model.

**Using matrix notation:**

- The covariates are $\mathbf{x} = (x_1, x_2)^T \in \mathbb{R}^2$ with no interaction or $\mathbf{x} = (x_1, x_2, x_1x_2)^T \in \mathbb{R}^3$ with an interaction.
- The parameters are $\beta = (\beta_1, \beta_2)^T$, or $\beta = (\beta_1, \beta_2, \beta_{12})^T$ if there is an interaction.
- The hazard rate is $h(t, x_1, x_2) = h_0(t) \exp(\beta^T \mathbf{x})$. 

3.1.8 A single binary covariate

- The covariate is $x \in \{0, 1\}$. If the set has alternative labels, relabel them 0 and 1.
- The parameter is $\beta \in \mathbb{R}$.
- The hazard rate is $h(t, x) = h_0(t) \exp(\beta x)$. In particular:
  \[
  h(t, 0) = h_0(t) \\
  h(t, 1) = h_0(t)e^\beta
  \]
  It is obvious that there are just two hazard rates and that for category 1 the hazard rate is $e^\beta$ times that for category 0.
- The hazard ratio for group 1 relative to group 0 is $e^\beta$.

Example: the effect of “race” on the effectiveness of the drug treatment. Individuals have been classified as “white” (or pinkish-grey as Steve Biko observed) and “other” (it is not clear what people with both European and non-European ancestry are classified as). Let us denote “white” as 0 and “other” as 1, and the “race” of individual $i$ as $f_i$.

The fitted hazard rate is:
\[
h(t, f) = h_0(t) \exp(-0.29f),
\]
that is, the hazard rate for “others” is 0.75 that of “whites”. The $p$-value is less than 1%, so there is strong evidence that the residential programme is better at treating “others” than “whites”.

3.1.9 A single categorical covariate or factor

- The covariate is $x \in \{c_0, c_1, \ldots, c_{K-1}\}$. We cannot simply relabel these $\{0, 1, \ldots, K-1\}$ as we did for two categories, as there is no reason why the hazard ratio for group $c_{i+1}$ relative to $c_i$ should be $e^\beta$. In R we can define these as factors. Alternatively, we can create dummy binary variables $C_{ik} = 1$ if $x_i = c_k$ and 0 otherwise. Note that we only create $K - 1$ of these, as $C_{i0} = 1 - \sum_{k=1}^{K-1} C_{ik}$.
• The parameters are $\beta_k \in \mathbb{R}$ for $k = 1, \ldots, K - 1$.

• The hazard rate is $h(t, x) = h_0(t) \exp(\sum_{k=1}^{K-1} \beta_k C_k)$. In particular:

\[
\begin{align*}
    h(t, c_0) &= h_0(t) \\
    h(t, c_1) &= h_0(t)e^{\beta_1} \\
    &\quad \cdots \\
    h(t, c_{K-1}) &= h_0(t)e^{\beta_{K-1}}
\end{align*}
\]

• The hazard ratio for group $c_k \neq 0$ relative to group $c_0$ is $e^{\beta_k}$. The hazard ratio for groups $c_k \neq 0$ and $c_j \neq 0$ is $\exp(\beta_k - \beta_j)$.

Example: effect of drug used on reversion to drug use. Each individual has been categorised according to heroin or cocaine use (particularly hard drugs). Category $c_0$ represents heroin and cocaine use, $c_1$ heroin only, $c_2$ cocaine only and $c_3$ is neither heroin nor cocaine. The fitted hazard function is:

\[
h(t, f) = h_0(t) \exp(0.078C_1 - 0.25C_2 - 0.16C_3),
\]

that is

\[
\begin{align*}
    h(t, c_0) &= h_0(t) \\
    h(t, c_1) &= h_0(t) \times 1.08 \\
    h(t, c_2) &= h_0(t) \times 0.78 \\
    h(t, c_3) &= h_0(t) \times 0.85.
\end{align*}
\]

Together these paint a confusing message: soft drug users appear to be at a lower risk of “re-offending” than those using both heroin and cocaine, while heroin-users appears more at risk than cocaine-users. However, the $p$-value of the model—i.e. assessing the null hypothesis that no predictor is associated with survival versus the alternative that at least one is—is only around 5%, and none of the parameters has an associated $p$-value of less than 5%. We would conclude that there is no strong evidence of an effect of the type of drug used.
3.1.10 A single categorical and a single continuous covariate

- The covariates are $x_1 \in \{0, 1\}$ and $x_2 \in \mathbb{R}$.
- The parameters are $(\beta_1, \beta_2) \in \mathbb{R}^2$ for the no-interaction model and $(\beta_1, \beta_2, \beta_{12}) \in \mathbb{R}^3$ for the interaction model.
- For the no-interaction model, the hazard rate is
  \[ h(t, x) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2). \]
  In particular:
  \[
  h(t, x_0 = 0, x_1) = h_0(t) e^{\beta_2 x_2} \\
  h(t, x_0 = 1, x_1) = h_0(t) e^{\beta_1 e^{\beta_2 x_2}}
  \]
- For the interaction model, the hazard rate is $h(t, x) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2)$. In particular:
  \[
  h(t, x_0 = 0, x_1) = h_0(t) e^{\beta_2 x_2} \\
  h(t, x_0 = 1, x_1) = h_0(t) e^{\beta_1 e^{(\beta_2 + \beta_{12}) x_2}}
  \]
- For the no-interaction model, the hazard ratio between the group with $x_1 = 1$ and the group with $x_1 = 0$ is $e^{\beta_1}$. The hazard ratio for a one-unit change in $x_2$ for either of the $x_1$ groups is $\exp(\beta_2)$.
- For the interaction model, the hazard ratio for a one-unit change in $x_2$ for the group with $x_1 = 0$ is $\exp(\beta_2)$, for the group with $x_1 = 1$ it is $\exp(\beta_2 + \beta_{12})$. The hazard ratio comparing the group with $x_1 = 0$ to that with $x_1 = 1$ depends on the value of the continuous covariate $x_2$.

Example: “race” and number of treatments for drug addiction. Each individual has been categorised as “white” ($x_1 = 0$) or “other” ($x_1 = 1$) and the number of previous treatments for drug addiction ($x_2$) has been recorded. The fitted hazard function for the no-interaction model is:

\[ h(t, f) = h_0(t) \exp(-0.26 x_1 + 0.027 x_2), \]
with tests of the hypotheses of the $\beta$s being equal to 0 giving $p = 1.7\%$ and $p = 0.05\%$ for $\beta_1$ and $\beta_2$, respectively.

The fitted hazard function for the interaction model is:

$$h(t, f) = h_0(t) \exp(-0.25x_1 + 0.027x_2 - 0.0014x_1x_2),$$

with an interaction term that cannot be statistically distinguished from 0 ($p = 0.96\%$). We thus prefer the no-interaction model.

### 3.1.11 Survival function for the Cox PHM

We already know that

$$S(t) = \exp \left \{ - \int_0^t h(\tau) \, d\tau \right \}$$

in general, so for the Cox PHM we have

$$S(t, x) = \exp \left \{ - \int_0^t h_0(\tau, \alpha) \exp(\beta^T x) \, d\tau \right \}$$

$$= \exp \left \{ - \int_0^t h_0(\tau, \alpha) \, d\tau \right \} \exp(\beta^T x)$$

$$= S_0(t, x, \alpha)^{\exp(\beta^T x)},$$

that is, a baseline survival function of unspecified form $S_0(t, x, \alpha)$ raised to the power $e^{\beta^T x}$.

### 3.2 Estimating parameters of Cox PHMs

If $\delta_i = 1$ if individual $i$ is uncensored and 0 if $i$ is right-censored for $i = 1, \ldots, m$, then we can write the likelihood function for a general model with some parameters $(\alpha, \beta)$ as

$$f(t|\alpha, \beta, x) = \prod_{i=1}^m h(t_i|\alpha, \beta, x)^{\delta_i} S(t_i|\alpha, \beta, x)$$
Specifically, for the Cox PHM we have
\[
f(t|\alpha, \beta, x) = \prod_{i=1}^{m} h(t_i|\alpha, \beta, x)^{\delta_i} S(t_i|\alpha, \beta, x)
\]
\[
= \prod_{i=1}^{m} h_0(t_i|\alpha)^{\delta_i} \exp(\beta^T x)^{\delta_i} S_0(t_i|\alpha)^{\exp(\beta^T x)}
\]
\[
\log f(t|\alpha, \beta, x) = \sum_{i=1}^{m} \delta_i \log \{h_0(t_i|\alpha)\} + \delta_i \beta^T x + \exp(\beta^T x) \log \{S_0(t_i|\alpha)\}.
\]

We cannot maximise this without specifying the form for the baseline hazard \(h_0(t_i|\alpha)\).

Instead we consider what is called the partial likelihood function. Here we define the risk set \(R\{t\}\) to be the set of all individuals \(i\) with \(t_i > t\), i.e. the people who haven’t died or been censored yet.

If survival times are continuous, we might expect that at any point in time, only one individual may instantaneously fail. However, because most observations are in fact slightly interval censored (e.g. they are recorded to the nearest whole month, say), it might not be the case in practice. This complicates things. We thus first consider the case when no individuals fail at the same time.

### 3.2.1 Partial likelihood for unique failure times

Throughout this chapter, let us use the notation \(\phi_i = \exp(\beta^T x_i)\), i.e. \(\phi_i\) is proportional to the hazard rate for individual \(i\) (the constant of proportionality being the baseline hazard function).

The partial likelihood for \(\beta\) is
\[
l_p(\beta, x) = \prod_{i=1}^{m} \left[ \frac{\phi_i}{\sum_{j \in R\{t_i\}} \phi_j} \right]^{\delta_i}.
\]
The $\delta_i$ power means that we only consider the contribution from death/failure times, not from the right-censored times. The numerator is proportional to the hazard for individual $i$, the one that has failed at time $t_i$. The denominator is proportional to the total hazard of all individuals (including $i$) that are at risk of failing at time $t_i$. So the fraction can be considered as the probability that it was $i$ and not some other individual that failed at the time $i$ failed.

There are two reasons why it is a partial likelihood:

- it is not the full likelihood for both $\alpha$ and $\beta$;
- it does not actually use the full data: the actual times events occur is not important, only their ranking. If individuals $i$, $j$ and $k$ fail at times 1, 2 and 3, respectively, this will give the same parameter estimates as if they had failed at times 100, 300, 1500, respectively.

It is thus less powerful than a fully parametric model. However, it requires fewer assumptions and so is more robust.

### 3.2.2 Partial likelihood for repeated failure times

The case when two or more individuals are recorded as failing at the same time is more complex. The exact partial likelihood for $\beta$ is considered last. First consider two approximations. The notation will be simpler (!) if we use the following notation:

- $t_{(i)}$ is the $i$th ordered unique failure time (so if four failures occur at times 1, 1, 3, 3, $t_{(1)} = 1$ and $t_{(2)} = 3$);
- $I$ is the total number of unique failure times;
- $D\{t\}$ is the set of individuals who fail at time $t$. 

Breslow’s method (Breslow, 1974):

\[ l_p(\beta, x) = \prod_{i=1}^{I} \frac{\prod_{j \in D\{t_{(i)}\}} \phi_j}{\left(\sum_{j \in R\{t_{(i)}\}} \phi_j\right)^{|D\{t_{(i)}\}|}}. \]

Note that \(|D\{t_{(i)}\}|\) is the number of individuals that fail at time \(t_{(i)}\).

Breslow’s method is the default for many statistical packages. But it is not the default for R. R uses Efron’s partial likelihood, as it is considered a closer approximation to the exact partial likelihood.

Efron’s method (Efron, 1977):

\[ l_p(\beta, x) = \prod_{i=1}^{I} \frac{\prod_{j \in D\{t_{(i)}\}} \phi_j}{\left(\sum_{j \in R\{t_{(i)}\}} \phi_j - \frac{k-1}{|D\{t_{(i)}\}|} \sum_{j \in D\{t_{(i)}\}} \phi_j\right)^{|D\{t_{(i)}\}|}}. \]

Exact method (Kalbfleisch and Prentice, 2002):

\[ l_p(\beta, x) = \prod_{i=1}^{I} \frac{\prod_{j \in D\{t_{(i)}\}} \phi_j}{\sum_{q \in Q_i} \Phi_q} \]

where \(Q_i\) is the set of all \(|D\{t_{(i)}\}|\)-tuples that could be selected from \(R\{t_{(i)}\}\) and \(\Phi_q\) is the product of \(\phi_j\) for all members \(j\) of \(|D\{t_{(i)}\}|\)-tuple \(q\).

An example

Suppose individuals labelled 1–5 are at risk at time \(t_{(i)}\), i.e. in \(R\{t_{(i)}\}\), and that of these, individuals 1–3 fail at time \(t_{(i)}\). Then Breslow’s method gives as the contribution from time \(t_{(i)}\) to the partial likelihood:

Efron’s method gives:
where
\[ v_a = (\phi_1 + \phi_2 + \phi_3 + \phi_4 + \phi_5) \]
\[ v_b = \left( \frac{2\phi_1}{3} + \frac{2\phi_2}{3} + \frac{2\phi_3}{3} + \phi_4 + \phi_5 \right) \]
\[ v_c = \left( \frac{\phi_1}{3} + \frac{\phi_2}{3} + \frac{\phi_3}{3} + \phi_4 + \phi_5 \right) \]

while the exact method gives:

You can see that the exact method quickly becomes computationally intensive when there are large numbers of ties. Note that in the absence of ties, all three reduce to the no-ties partial likelihood.

### 3.3 Estimating the parameters numerically

The three versions of the partial likelihood function described earlier are fairly difficult to fit analytically. Luckily, a computer can do it for you. R uses the Newton–Raphson method to estimate the parameters. This is a method that often but not always converges to the desired maximum likelihood estimates. Because it does not always succeed, it is worth having an overview of the method.

The Newton–Raphson method is a deterministic, iterative procedure. It is deterministic because there is no element of randomness in the search for the optimum (in contrast to stochastic procedures such as simulated annealing and cross entropy). It is iterative because it consists of a series of iterations, with the estimate (hopefully) getting better at each iteration.

In general, if we have a parameter vector \( \theta \) of dimension \( p \) and wish to find \( \hat{\theta} \) which maximises the log-likelihood function \( l(\theta) \), the algorithm is:

1. \[ \text{ } \]
2.

3.

4.

5.

Here we have used the notation

- $\theta^{(k)}$ is the value of the parameters at iteration $k$ of the routine;
- the score function is
  \[ U(\theta) = \left( \frac{\partial l(\theta)}{\partial \theta_1}, \ldots, \frac{\partial l(\theta)}{\partial \theta_p} \right); \]
- the information matrix is
  \[ I(\theta) = - \begin{pmatrix} \frac{\partial^2 l(\theta)}{\partial \theta_1^2} & \cdots & \frac{\partial^2 l(\theta)}{\partial \theta_1 \partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\theta)}{\partial \theta_p \partial \theta_1} & \cdots & \frac{\partial^2 l(\theta)}{\partial \theta_p^2} \end{pmatrix}; \]
- and $\theta_q$ is the $q$th element of $\theta$.

Note:

- although the choice of $\theta^{(0)}$ is arbitrary, the further it is from $\hat{\theta}$, the less likely the algorithm is to converge to $\hat{\theta}$;
- the log likelihood $l(\theta)$ may be replaced by the partial log likelihood $l_p(\theta)$;
- in the Cox PHM case, we may write $\beta$ instead of $\theta$. 
Let’s try an example. Consider a continuous covariate $x_i$, equal to the body mass index of individual $i$. We have 9 individuals who have suffered a heart attack; $t_i$ is the time of death of $i$ in days following the attack. None of these 9 are censored. The data are:

<table>
<thead>
<tr>
<th>$i$</th>
<th>$t_i$</th>
<th>$x_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>31.4</td>
</tr>
<tr>
<td>2</td>
<td>98</td>
<td>21.5</td>
</tr>
<tr>
<td>3</td>
<td>189</td>
<td>27.1</td>
</tr>
<tr>
<td>4</td>
<td>374</td>
<td>22.7</td>
</tr>
<tr>
<td>5</td>
<td>1002</td>
<td>35.7</td>
</tr>
<tr>
<td>6</td>
<td>1205</td>
<td>30.7</td>
</tr>
<tr>
<td>7</td>
<td>2065</td>
<td>26.5</td>
</tr>
<tr>
<td>8</td>
<td>2201</td>
<td>28.3</td>
</tr>
<tr>
<td>9</td>
<td>2421</td>
<td>27.9</td>
</tr>
</tbody>
</table>

Let us fit the model $h(t, x_i) = h_0(t, \alpha)e^{\beta x_i}$ to the data using maximum likelihood. There are no ties in the survival times, so we may use the simplest partial log likelihood function (actually all three complex ones give this as the special case when there are no ties):

$$l_p(\beta) = \beta \sum_{i=1}^{9} x_i - \sum_{i=1}^{9} \log \left( \sum_{j \in R\{t_i\}} e^{\beta x_j} \right)$$

$$\frac{d}{d\beta} l_p(\beta) = \sum_{i=1}^{9} x_i - \sum_{i=1}^{9} \frac{\sum_{j \in R\{t_i\}} x_j e^{\beta x_j}}{\sum_{j \in R\{t_i\}} e^{\beta x_j}}$$

$$\frac{d^2}{d\beta^2} l_p(\beta) = -\sum_{i=1}^{9} \frac{A}{\left( \sum_{j \in R\{t_i\}} e^{\beta x_j} \right)^2}$$

where $A = \left( \sum_{j \in R\{t_i\}} x_j^2 e^{\beta x_j} \right) \left( \sum_{j \in R\{t_i\}} e^{\beta x_j} \right) - \left( \sum_{j \in R\{t_i\}} x_j e^{\beta x_j} \right) \left( \sum_{j \in R\{t_i\}} x_j^2 e^{\beta x_j} \right)$.
These look pretty horrible (more than one covariate looks worse!) but the important thing is that we can evaluate them for a particular value of $\beta$ with no difficulty. If we let $U(\beta) = d l_p(\beta)/d\beta$ and $I(\beta) = -d^2 l_p(\beta)/d\beta^2$, the Newton–Raphson update step is simply

$$\beta^{(k+1)} = \beta^{(k)} = U(\beta^{(k)})/I(\beta^{(k)}).$$

Let our first guess at the MLE be

$$\beta^{(0)} = 0.$$

Then $u(0) = -2.51$ and $I(0) = 77.13$, so our new guess at the MLE is

$$\beta^{(1)} = 0 - 2.51/77.13 = -0.0326.$$

The next iteration gives

$$\beta^{(2)} = -0.0326 - 0.069/72.83 = -0.0335.$$

The next iteration gives

$$\beta^{(3)} = -0.0335 - 0.000061/72.70 = -0.0335 = \beta^{(2)}$$

to 3 significant figures. We can stop here if this level of accuracy is sufficient.

This method works well if the starting value is sufficiently close to the target value. But if not, it can result in massive jumps far away from the target value. To reduce the risk of this happening, we can change the update step to

$$I(\theta^{(k)}) (\theta^{(k+1)} - \theta^{(k)}) = \xi U(\theta^{(k)})$$

where $\xi < 1$ acts as a brake limiting the size of the jump. This increases the number of iterations required to reach the target value.
3.4 Estimating the parameters in R

You don’t need to know the details of how software packages fit this model unless something goes wrong. In R, the Cox PHM can be fit to data quite easily using the `coxph` command. This requires a *formula object* of form `Surv() ~ covariates` (c.f. last chapter). It has optional arguments of form:

```r
, data = aml  # a data frame called aml
, subset = 1:100  # use only the first 100 data
, init = c(0.01,0.01)  # use 0.01 0.01 as initial values for
  # beta1 beta2 instead of 0 0
, method = 'efron'  # use Efron’s method to deal with ties
    # (Efron’s is the default)
, method = 'breslow'  # use Breslow’s method to deal with ties
, method = 'exact'  # use the exact method to deal with ties
```

For example, we might load the full data set on survival following heart attacks (from the Worcester heart attack study, see Hosmer et al. (2008) and fit the Cox PHM for the effect of body mass index (BMI) as follows:

```r
library(survival)
cols=c('id','age','sex','c4','c5','c6','bmi','history','c9','c10','c11',
  'c12','c13','c14','c15','c16','c17','c18','c19','c20','t','delta')
attach(read.table('http://www.umass.edu/statdata/statdata/data/whas500.dat'
  ,col.names=cols))
phm.bmi = coxph(Surv(t,delta)~bmi)
summary(phm.bmi)
```

So doing yields the following output:
Call:
coxph(formula = Surv(t, delta) ~ bmi)

<table>
<thead>
<tr>
<th>coef</th>
<th>exp(coef)</th>
<th>se(coef)</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>bmi</td>
<td>-0.0985</td>
<td>0.906</td>
<td>0.0148</td>
<td>-6.68</td>
</tr>
</tbody>
</table>

exp(coef) exp(-coef) lower .95 upper .95
bmi 0.906 1.10 0.88 0.933

Rsquare= 0.092  (max possible= 0.993 )
Likelihood ratio test= 48.3 on 1 df,  p=3.6e-12
Wald test = 44.6 on 1 df,  p=2.46e-11
Score (logrank) test = 44.3 on 1 df,  p=2.81e-11

So our fitted model is

$$h(t, x_i) = h_0(t, \alpha) \exp(-0.0985 x_i)$$

where $x_i$ is the BMI of individual $i$. The exp(coef) term indicates that the hazard ratio for a one unit increase in BMI is around 90%, so that increasing your weight such that your BMI goes up by one leads to a reduction in risk of death of 10% among heart attack survivors.

Let us repeat the analysis using the Breslow partial likelihood. We can time how long each routine takes using the \texttt{system.time()} command. The Breslow and Efron methods took less than one second, while the exact method took over one and a half hours before I gave up. Here are the output of the Breslow method:

Call:
coxph(formula = Surv(t, delta) ~ bmi, method = "breslow")

<table>
<thead>
<tr>
<th>coef</th>
<th>exp(coef)</th>
<th>se(coef)</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>bmi</td>
<td>-0.0983</td>
<td>0.906</td>
<td>0.0147</td>
<td>-6.67</td>
</tr>
</tbody>
</table>
\[
\begin{array}{cccc}
\exp(\text{coef}) & \exp(-\text{coef}) & \text{lower .95} & \text{upper .95} \\
bmi & 0.906 & 1.10 & 0.88 & 0.933 \\
\end{array}
\]

\[
\text{Rsquare} = 0.092 \quad \text{(max possible} = 0.993 \text{)} \\
\text{Likelihood ratio test} = 48.2 \text{ on 1 df}, \quad p=3.92\text{e-12} \\
\text{Wald test} = 44.4 \text{ on 1 df}, \quad p=2.64\text{e-11} \\
\text{Score (logrank) test} = 44.2 \text{ on 1 df}, \quad p=3.02\text{e-11} \\
\]

Here, the approximations are very similar, but usually Efron’s approximation is closer to the exact estimate.

Among the other output we can see that R provides us with the standard error of \( \beta \). This is obtained by noting that

\[
\hat{\beta} \sim N(\beta, E\{I(\beta)\}^{-1})
\]

asymptotically. Actually, what R provides is an estimated standard error, obtained by replacing the expected information \( E\{I(\beta)\} \) by the observed information \( I(\hat{\beta}) \).

We can use this estimated standard error and approximate distribution to obtain confidence intervals for \( \beta \). If \( z_{1-\alpha/2} \) is the \((1 - \alpha/2)\)ile of the standard normal distribution, the \( 1 - \alpha \) confidence interval is

\[
\hat{\beta} \pm z_{1-\alpha/2} \frac{1}{\sqrt{I(\hat{\beta})}} \quad \text{i.e.} \quad \hat{\beta} \pm z_{1-\alpha/2} \hat{se}(\hat{\beta}).
\]

For instance, using the Efron output, we have the following 95% confidence interval for \( \beta \) of

\[-0.0985 \pm 1.96 \times 0.0148 = (-0.13, -0.07).\]

It can be seen from the output that R also calculates some \( p \)-values. In the next section, we consider how these are found.
3.5 Hypothesis testing for the PHM

There are two tests that will be very useful in testing the hypothesis that one or more covariates have no effect. These are the Wald and the (partial) likelihood ratio tests. (A further test exists: the score test, but it doesn’t really add anything to the others.) Both test:

\[
H_0 : \beta = 0 \\
H_1 : \beta \neq 0
\]

for the model \( h(t, x_i) = h_0(t, \alpha) e^{\beta x_i} \). For models with multiple parameters, it is most often convenient to use the Wald test for one parameter at a time, as it requires the model be fitted once only. When fitting different, nested models, the likelihood ratio test is most convenient. Hosmer et al. (2008) suggest that the likelihood ratio test is preferable in situations in which all three tests may be applied.

- For a test of a single parameter being equal to 0, the Wald test statistic is

\[
z^2 = \frac{\hat{\beta}^2}{\hat{V}(\hat{\beta})}.
\]

If \( H_0 \) is true, \( z^2 \sim \chi^2_1 \) (or, equivalently, \( z \sim N(0, 1) \)). Large values of \( z^2 \) support the alternative hypothesis. For multivariate models, a version of the Wald test exists, which comes from a \( \chi^2 \) distribution with more degrees of freedom, but you will rarely need this.

- The likelihood ratio test statistic for the hypothesis that a single parameter is equal to zero is

\[
G = 2[l_p(\hat{\beta}) - l_p(0)].
\]

It too should be \( G \sim \chi^2_1 \) if \( H_0 \) is true. For tests of multiple parameters being equal to zero, the degrees of freedom increase as explained below.

The three tests (Wald, score, LRT) give different \( p \)-values in general.
Actually, we may make the null hypothesis more general. Let us consider a $p$-dimensional parameter vector $\beta$, where without loss of generality we wish to test the hypothesis that the first $1 \leq q \leq p$ elements of $\beta$ are equal to some specified values, $\beta^*_j$ say, for $j = 0, \ldots, q$. The remaining $p - q$ elements are free parameters. The alternative hypothesis is that at least one of these $q$ parameters is not equal to the hypothesised value.

We perform the test by fitting two nested models. The general model lets all $p$ parameters be estimated by maximum likelihood. The particular model fixes the first $q$ parameters at their hypothesised values but lets the remaining $p - q$ be estimated by maximum likelihood. The parameter space for the general model contains the lower dimensional parameter space for the particular model, i.e. the particular model is nested within the general model.

Let $l_p(\hat{\beta})$ be the value of the (partial) log likelihood at the maximum likelihood estimates $\hat{\beta}$ of the general model.

Let $l_p(\beta^*_{1:q}, \hat{\beta}_{q+1:p})$ be the value of the (partial) log likelihood at the maximum likelihood estimate of the $p - q$ free parameters conditioned on the $q$ fixed parameters of the particular model.

If $H_0$ is true, $G = 2[l_p(\hat{\beta}) - l_p(\beta^*_{1:q}, \hat{\beta}_{q+1:p})]$ has a chi-squared distribution with $q$ degrees of freedom.

Note:

- Usually you use this in selecting between two models, one of which has an additional parameter, perhaps for an interaction, which you wonder might equal zero.
- The special case given above may be recovered when $p = q = 1$ and $\beta^*_1 = 0$.
- This test may be used for log likelihoods in other settings as well as for the partial log likelihoods here.
- It allows us to test whether the hazard ratio is some particular constant, not just $1 = e^0$. 
• When R provides output

\[
\begin{align*}
\text{Likelihood ratio test} &= 48.3 \text{ on } 1 \text{ df, } p=3.6\times10^{-12} \\
\text{Wald test} &= 44.6 \text{ on } 1 \text{ df, } p=2.46\times10^{-11} \\
\text{Score (logrank) test} &= 44.3 \text{ on } 1 \text{ df, } p=2.81\times10^{-11}
\end{align*}
\]

it is testing the hypothesis that all parameters are equal to their hypothesised values, here 0. You should ignore the Wald and score test results.

### 3.5.1 An example

Let us return to the Worcester heart attack data. There are four covariates that we have labelled with names. These are: body mass index (let us denote this \( x_i^b \) for individual \( i \) henceforth), age at heart attack (\( x_i^a \)), sex (\( x_i^s = 0 \) for male and 1 for female [the other way around would have allowed the number of y-chromosomes to be an aide-mémoire]) and history of cardiovascular disease (\( x_i^h = 1 \) if \( i \) has a history and 0 otherwise).

In the next chapter we consider the best way to develop a model with multiple potential predictors. For now, let us approach the data naïvely by first fitting each term individually.

\[
\begin{align*}
\text{phm.bmi} &= \text{coxph(Surv(t,delta)~bmi)} \\
\text{phm.age} &= \text{coxph(Surv(t,delta)~age)} \\
\text{phm.sex} &= \text{coxph(Surv(t,delta)~sex)} \\
\text{phm.history} &= \text{coxph(Surv(t,delta)~history)} \\
\text{summary(phm.bmi)} \\
\text{summary(phm.age)} \\
\text{summary(phm.sex)} \\
\text{summary(phm.history)}
\end{align*}
\]

These yield the following output:
Body mass index clearly is associated with an effect on survival (not necessarily causal).

Age, too, affects survival, with older heart attack patients at more risk of death than younger (though not necessarily more than would be expected among non-heart attack patients of differing ages).
Sex also has an impact upon survival. Surprisingly, females are at a higher risk of death than males.

In contrast, having a history of heart problems does not increase the risk significantly of death following a heart attack (though it is probably highly influential in predicting the heart attack in the first place).
Since BMI, age and sex all seem to influence survival, a natural second step would be to fit a hazard function incorporating all three:

\[ h(t, x_i^a, x_i^b, x_i^s) = h(0, \alpha) \exp(\beta_a x_i^a + \beta_b x_i^b + \beta_s x_i^s). \]

Doing this in R gives the following output:

\texttt{Call:}
\texttt{coxph(formula = Surv(t, delta) ~ bmi + age + sex)}

\begin{verbatim}
 n= 500

 coef exp(coef) se(coef) z p
 bmi -0.0421 0.959 0.0154 -2.73 0.0064
 age 0.0608 1.063 0.0065 9.35 0.0000
 sex -0.0931 0.911 0.1411 -0.66 0.5100

 exp(coef) exp(-coef) lower .95 upper .95
 bmi 0.959 1.043 0.930 0.988
 age 1.063 0.941 1.049 1.076
 sex 0.911 1.098 0.691 1.201

 Rsquare= 0.259  (max possible= 0.993 )
 Likelihood ratio test= 150 on 3 df,  p=0
 Wald test = 129 on 3 df,  p=0
 Score (logrank) test = 135 on 3 df,  p=0
\end{verbatim}

Sex is no longer predictive of survival. It seems that sex is associated with one of the other predictors. Checking these, we find that the mean age of men in the study is 67 years, but of women 75. Similarly the mean BMI for males was 27.3, but of females, 25.6. Thus being male is associated with being younger and heavier, both of which are associated with longer survival times. We therefore drop sex from the model. Since both other factors are significant on their own, we also investigate whether there is an interaction between them, i.e. we fit the model:

\[ h(t, x_i^a, x_i^b) = h(0, \alpha) \exp(\beta_a x_i^a + \beta_b x_i^b). \]
CHAPTER 3. COX PHM

Call:
coxph(formula = Surv(t, delta) ~ bmi * age)

n= 500

<table>
<thead>
<tr>
<th></th>
<th>coef</th>
<th>exp(coef)</th>
<th>se(coef)</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>bmi</td>
<td>-0.030818</td>
<td>0.97</td>
<td>0.09168</td>
<td>-0.336</td>
<td>0.740</td>
</tr>
<tr>
<td>age</td>
<td>0.063526</td>
<td>1.07</td>
<td>0.03053</td>
<td>2.081</td>
<td>0.037</td>
</tr>
<tr>
<td>bmi:age</td>
<td>-0.000134</td>
<td>1.00</td>
<td>0.00117</td>
<td>-0.115</td>
<td>0.910</td>
</tr>
</tbody>
</table>

exp(coef) exp(-coef) lower .95 upper .95

<table>
<thead>
<tr>
<th></th>
<th>exp(coef)</th>
<th>exp(-coef)</th>
<th>lower .95</th>
<th>upper .95</th>
</tr>
</thead>
<tbody>
<tr>
<td>bmi</td>
<td>0.97</td>
<td>1.031</td>
<td>0.810</td>
<td>1.16</td>
</tr>
<tr>
<td>age</td>
<td>1.07</td>
<td>0.938</td>
<td>1.004</td>
<td>1.13</td>
</tr>
<tr>
<td>bmi:age</td>
<td>1.00</td>
<td>1.000</td>
<td>0.998</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Rsquare= 0.259 (max possible= 0.993 )

Likelihood ratio test= 150 on 3 df, p=0
Wald test = 129 on 3 df, p=0
Score (logrank) test = 154 on 3 df, p=0

The fact that the p-value for an interaction is large tells us to dump this model and return to the no-interaction model:

\[ h(t, x_i^a, x_i^b) = h(0, \alpha) \exp(\beta_a x_i^a + \beta_b x_i^b). \]

Call:
coxph(formula = Surv(t, delta) ~ bmi + age)

n= 500

<table>
<thead>
<tr>
<th></th>
<th>coef</th>
<th>exp(coef)</th>
<th>se(coef)</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>bmi</td>
<td>-0.0412</td>
<td>0.96</td>
<td>0.01529</td>
<td>-2.69</td>
<td>0.007</td>
</tr>
<tr>
<td>age</td>
<td>0.0601</td>
<td>1.06</td>
<td>0.00643</td>
<td>9.34</td>
<td>0.000</td>
</tr>
</tbody>
</table>

exp(coef) exp(-coef) lower .95 upper .95

<table>
<thead>
<tr>
<th></th>
<th>exp(coef)</th>
<th>exp(-coef)</th>
<th>lower .95</th>
<th>upper .95</th>
</tr>
</thead>
<tbody>
<tr>
<td>bmi</td>
<td>0.96</td>
<td>1.042</td>
<td>0.931</td>
<td>0.989</td>
</tr>
<tr>
<td>age</td>
<td>1.06</td>
<td>0.942</td>
<td>1.049</td>
<td>1.075</td>
</tr>
</tbody>
</table>

Rsquare= 0.259 (max possible= 0.993 )
Likelihood ratio test = 150 on 2 df,  p=0
Wald test = 128 on 2 df,  p=0
Score (logrank) test = 135 on 2 df,  p=0

From this output, we can see that the hazard ratio for a one-unit increase in BMI is 0.96, with 95% confidence interval (0.93,0.99) after adjusting for age. The hazard ratio for a one-year increase in age is 1.06 (1.05,1.08), adjusting for BMI. Our final model is thus:

\[
h(t, x_i^a, x_i^b) = h(0, \alpha) \exp(0.060x_i^a - 0.041x_i^b).
\]
3.6 Estimating the baseline hazard function

The Cox PHM is *semi-parametric* method of estimation. We specify a model for the effect of the covariates but we don’t specify a model for the baseline hazard function. When we used the Kaplan–Meier method we also did not specify a model for the survival function. Since the hazard and survival functions are intimately linked, we can thus adapt the Kaplan–Meier method to estimate the baseline hazard function.

The estimate for the baseline hazard function at the time $t(i)$ of the $i$th event is

$$\hat{h}_0(t(i)) = \frac{d(i)}{\sum_{j \in R\{t(i)\}} \exp(\hat{\beta}^T x_j)}$$

where

- $d(i)$ is the number of deaths at that time; and
- $R\{t(i)\}$ is the set of individuals that could die at that time.

3.7 Estimating the baseline survival function

Recall that $S(t) = \exp(- \int_0^t h(\tau) \, d\tau)$. Here, our estimate of the hazard function is a discrete approximation to a continuous function, namely we use $\hat{h}_0(t(i))$ as above to estimate $\int_{t(i-1)}^{t(i)} h_0(\tau) \, d\tau$. Thus our estimate of the baseline survival function is

$$\hat{S}_0(t(i)) = \exp \left[- \sum_{j \leq i} \hat{h}_0(t(j)) \right].$$
3.8 Plotting adjusted survival functions in R

Unfortunately, it is not a simple task to plot survival functions in R accounting for a Cox proportional hazards model. Firstly we need to decide what exactly we wish to plot. Consider the following motivating example.

Drug users have been allocated to one of two residential treatments to try to reduce the risk of them reoffending: being randomly allocated to either a long or a short stay. A variety of further covariates were recorded. One such was the number of previous treatments the addict had had for his or her addiction. Letting $x^L_j = 1$ if individual $j$ was allocated to the long treatment and 0 if allocated to the short, and $x^N_j$ be the number of previous drug treatments $j$ received (here in the range $[0, 40]$ albeit with some missing values), we fit the following model:

$$h(t, x^L_j, x^N_j) = h_0(t) \exp(\beta_L x^L_j + \beta_N x^N_j)$$

(the interaction term being non-significant). Note we have dropped the parameters $\alpha$ from the baseline hazard as henceforth we shall be estimating this function non-parametrically.

Here is the R output from fitting this model.

```R
Call: coxph(formula = S ~ ndrugtx + treat)
n=611 (17 observations deleted due to missingness)
             coef exp(coef) se(coef)      z     p
ndrugtx  0.0300  1.0300    0.0075  3.97 7.1e-05
    treat -0.2220  0.8010    0.0900 -2.47 1.4e-02

exp(coef) exp(-coef) lower .95 upper .95
ndrugtx  1.0300    0.97  1.0150   1.0460
    treat  0.8010    1.25  0.6710   0.9560
```
Both covariates influence survival. Having a long residential treatment decreases the risk of reoffending by about 20%.

What do we want to plot? We can only plot $\hat{S}(t, x^L, x^N)$ versus $t$ for some specified values of $x^L$ and $x^N$. For instance, we might plot $\hat{S}(t, 0, 0)$ and $S(t, 1, 0)$ versus $t$, that is, the survival function of an addict with no previous treatments on a short or a long residential stay, respectively. Our reader would more likely wish to see some kind of average over the non-plotted covariates: for example, a plot of the survival function of an average addict on a short or a long residential stay.

However, the definition of “average” is not clear cut. It could be the average (number of existing treatments) over the whole sample; it could be the average over the subset of the sample that had that particular length of stay. For a covariate that has been randomly allocated, there is little real difference between these (unless you have been unlucky in the randomisation); however for non-randomised covariates, this is not the case. In particular, the two covariates may be correlated. This is a particular problem in some epidemiological studies, where we cannot randomly allocate covariates such as history of drug treatments.

The best solution is to control for the non-plotted covariate by taking the mean value of the covariate for the subset of the sample that you are plotting. So for our example, we would take the mean value of $x^N$ in the group with $x^L = 0$ and use that in our plot for $\hat{S}(t, x^L = 0)$ versus $t$, and vice versa for the $x^L = 1$ plot. You should explain clearly in the description of the methods how you have done so when writing a report of it.

Warning! This is a good place for a warning. Even if a covariate is associated with a change in survival, it need not be the cause of that change. For example, the number of previous drug treatments might be correlated but not the cause of an increased risk of reoffending. In this case it could well be correlated with another covariate that really is the cause. One of the strongest indications that an observed correlation is the result of a cause and effect relationship is when we have randomised the allocation of that covariate, for instance treatment of long or short residential stays. In observational studies, one usually tries to adjust for potential confounders by including them in the regression equation.
Suppose that we have decided on the covariates we wish to use in the plot, and the particular values are $x$. Then we first must estimate the baseline survival function $S_0(t)$ and then raise this to the power of $\exp(\beta^T x)$. To find the baseline survival function, adapt the following commands. These exploit the fact that the estimates of the hazard are provided as output in the `coxph.detail()` function.

```r
phm = coxph(Surv(t,delta)~x)
d.phm = coxph.detail(phm)
times = c(0,d.phm$t)
h0 = c(0,d.phm$hazard)
S0 = exp(-cumsum(h0))
```

Here $x$ is a covariate (or vector of covariates), $t$ is the vector of event times and $\delta$ indicates if death occurs.

One annoying thing with the implementation of this in R is that the function `coxph.detail` returns not the estimate of the baseline hazard, but rather the estimate of the hazard for an average individual, i.e. with all covariates evaluated at their average over the sample. Thus, to evaluate the estimate at a particular value $x$, we must subtract from this value the mean $\bar{x}$ in the following way:

```r
beta = phm$coef
meanx = c(mean(x,na.rm=T)) #a vector if more than one covariate
x = c(0)-meanx
Sx = S0 ^ exp(t(beta) %*% x)
```

Now suppose that we have two covariates: $x1$ which may be 0 or 1 and $x2$ which is a real; we wish to plot $x1$ controlled for $x2$. The following commands will do the trick.
#no interaction model:

```r
phm = coxph(Surv(t,delta)~x1+x2)
d.phm = coxph.detail(phm)
times = c(0,d.phm$t)
h0 = c(0,d.phm$hazard)
S0 = exp(-cumsum(h0))
beta = phm$coef
meanx = c(mean(x1),mean(x2))
x_A = c(0,mean(subset(x2,x1==0)))-meanx
Sx_A = S0 ^ exp(t(beta) %*% x_A)
x_B = c(1,mean(subset(x2,x1==1)))-meanx
Sx_B = S0 ^ exp(t(beta) %*% x_B)
xlb='t'; ylb=expression(hat(S)(t))
plot(times,Sx_A,type='s',xlab=xlb,ylab=ylb,ylim=0:1)
lines(times,Sx_B,col=2,type='s')
```

Adapting this for our addicts example will yield a plot like this. For comparison, the Kaplan–Meier estimates are also shown (as dashed lines).
We could do something similar to the previous section in order to plot the adjusted hazard function. However, this isn’t very interesting, as it is strongly affected by the discrete nature of the method of estimation, as shown in the following figure.
3.9 Confidence intervals for functions of one or more parameters

Thus far we have fleetingly touched upon confidence intervals for a parameter \( \beta \) of the Cox proportional hazards model. We found out already that the asymptotic distribution of the maximum (partial) likelihood estimates of a parameter vector \( \beta \) was

\[
\hat{\beta} \sim N(\beta, \text{E}(I(\beta))^{-1})
\]

in which we can replace the expected information matrix by an estimate, namely the observed information, thus:

\[
\hat{\beta} \sim N(\beta, \text{I}(\hat{\beta})^{-1}).
\]

If all we wish to do is to make a confidence interval for \( \beta_i \), i.e. for one element of this parameter vector, we can just use the output R gives us, i.e. the standard error of \( \beta_i \). However, we may want more. We may wish to create confidence intervals for the hazard ratio between two categories, or between two individuals with multiple different covariates. This is not automatically available through the R output, but is fairly easy to obtain from what R does give us.

We first go over the properties of the multivariate normal distribution and the covariance function.

3.9.1 The multivariate normal distribution

If \( \mathbf{x} \sim N(\mu, \Sigma) \) then any subset of the rows of \( \mathbf{x} \) also has a Normal distribution with corresponding rows of \( \mu \) and rows and columns of \( \Sigma \). This means that, for example, if we have two covariates we are interested in, we can discard the parameters for all the other covariates and end up with a bivariate Normal distribution for the joint distribution of the parameters for those two covariates of interest.
For example,
\[
\begin{pmatrix}
\hat{\beta}_1 \\
\hat{\beta}_2 \\
\hat{\beta}_3
\end{pmatrix}
\sim N
\begin{pmatrix}
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
\beta_3
\end{pmatrix},
\begin{pmatrix}
\sigma_{11} & \sigma_{12} & \sigma_{13} \\
\sigma_{21} & \sigma_{22} & \sigma_{23} \\
\sigma_{31} & \sigma_{32} & \sigma_{33}
\end{pmatrix}
\end{pmatrix}
\Rightarrow
\begin{pmatrix}
\hat{\beta}_1 \\
\hat{\beta}_3
\end{pmatrix}
\sim N
\begin{pmatrix}
\begin{pmatrix}
\beta_1 \\
\beta_3
\end{pmatrix},
\begin{pmatrix}
\sigma_{11} & \sigma_{13} \\
\sigma_{31} & \sigma_{33}
\end{pmatrix}
\end{pmatrix}
\]

### 3.9.2 Properties of the covariance

We need to refresh ourselves with the properties of the covariance function. You should know that if we have two independent variables $X$ and $Y$, then $V(X + Y) = V(X) + V(Y)$. In general, variables are not automatically independent. It is certainly the case that the MLEs of our $\beta$s will not be independent. We must therefore generalise this to $V(X + Y) = V(X) + V(Y) + 2C(X, Y)$, where $C(X, Y)$ is the covariance of $X$ and $Y$. The covariance has the following properties:

- $C(X, Y) = C(Y, X)$
- $C(X, X) = V(X)$
- $C(aX + b, cY + d) = acC(X, Y)$

### 3.9.3 Putting these together

From these, we can work out that
\[
V(\hat{\beta}_i - \hat{\beta}_j) = V(\hat{\beta}_i) + V(\hat{\beta}_j) - 2C(\hat{\beta}_i, \hat{\beta}_j).
\]

Thus, a 95% CI for $\beta_i - \beta_j$ is
\[
\hat{\beta}_i - \hat{\beta}_j \pm 1.96 \sqrt{\sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}}.
\]

R will give us the estimates of the coefficients and variance–covariance matrix. To obtain these, fit the model using the `coxph`, and call the `var` or `coefficients`. For example,
phm.cell=coxph(Surv(t,delta)~factor(Cell))
phm.cell$var
phm.cell$coefficients

For example, we may get the following output

> phm.cell$coefficients
  factor(Cell)large factor(Cell)small factor(Cell)squamous
-1.0012532 0.1464599 -0.7711077
> phm.cell$var
[,1]   [,2]   [,3]
[1,] 0.06426602 0.02031726 0.02559262
[2,] 0.02031726 0.06214750 0.02151569
[3,] 0.02559262 0.02151569 0.06381066

from fitting the Cox PHM to the data on veterans’ survival following lung cancer diagnosis considered in Tutorial 3b. This leads to the following 95% confidence interval for the difference between squamous and large cells:

\[-0.77 - 1.00 \pm 1.96 \sqrt{0.0642 + 0.0638 - 2 \times 0.0256}\]

i.e. \((-0.31, 0.77)\).

A 95% confidence interval for the hazard ratio comparing squamous to large cells is obtained by exponentiating the end points, thus \((e^{-0.31}, e^{0.77}) = (0.73, 2.17)\).

We can generalise this approach to obtain confidence intervals for any linear combinations of the regression coefficients. For example, if we have two covariates and wish to compare the hazard ratio for two individuals with values of the covariates equal to \(x^a = (x^a_1, x^a_2)\) and \(x^b = (x^b_1, x^b_2)\), the variance of \(\hat{\beta}^T x^a - \hat{\beta}^T x^b\) is

\[
V(\hat{\beta}^T x^a - \hat{\beta}^T x^b) = V\{\hat{\beta}_1(x^a_1 - x^b_1) + \hat{\beta}_2(x^a_2 - x^b_2)\} \\
= (x^a_1 - x^b_1)^2 V(\hat{\beta}_1) + (x^a_2 - x^b_2)^2 V(\hat{\beta}_2) \\
+ 2C\{\hat{\beta}_1(x^a_1 - x^b_1), \hat{\beta}_2(x^a_2 - x^b_2)\} \\
= (x^a_1 - x^b_1)^2 V(\hat{\beta}_1) + (x^a_2 - x^b_2)^2 V(\hat{\beta}_2) \\
+ 2(x^a_1 - x^b_1)(x^a_2 - x^b_2)C\{\hat{\beta}_1, \hat{\beta}_2\}.
\]