MAS3301 / MAS8311 Biostatistics
Part II: Survival

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 2, 2009-10

11 Incorporating covariates: proportional hazards models

Up to this stage of the course we have generally not had information for each individual other than the survival time and censoring status ie. we have not considered information such as the weight, age, or smoking status of individuals, for example. These are referred to as *covariates* or *explanatory variables*.

For the remainder of the course we will be considering how to incorporate covariates into our analysis. In this way, for example, individuals with different ages / weights / smoking status will be modelled via different (but related) life time distributions. In addition, we will be able to make decisions about the life time of an individual based on their covariate values (ie. ages, weight, etc).

The are essentially two simple ways to incorporate covariates:

- Proportional hazards models (which are very commonly used).
- Accelerated life models (which are used more rarely).

11.1 Proportional hazards models

Suppose we have individuals $i = 1, \ldots, n$ in a study, and for each one we record a set of covariates $x_{i1}, x_{i2}, \ldots, x_{ip}$. For example x_{i1} could be the weight of individual i and x_{i2} the body mass index (BMI). We denote the hazard function of individual i by $h_i(t)$. Note that previously we have generally had a single hazard function that we used for all individuals. In a proportional hazards model we assume that for any two individuals i, j the hazards are related by

$$h_i(t) = \psi_{ij} \times h_j(t)$$

where ψ_{ij} is a constant that *does not depend on t*. Usually this assumption is written slightly differently (but equivalently) as

$$h_i(t) = \phi_i \times h_0(t)$$

for a constant ϕ_i where $h_0(t)$ is the *baseline* hazard function. The constants ϕ_i depend on the covariates. In other words, every individual has a hazard function that is a constant multiple of the underlying baseline hazard, where the constant of proportionality depends on the covariates.

Example:

Suppose we have a very small study with just 4 individuals and 2 covariates:

Individual	Smoker?	Weight
1	Yes	70kg
2	No	65kg
3	Yes	82kg
4	Yes	87kg

Then, for example, we might assume a Weibull baseline hazard

$$h_0(t) = \lambda\gamma t^{\gamma-1}$$

and then a proportional hazards model, so that

$$h_i(t) = \phi_i \times h_0(t) = \phi_i \times \lambda\gamma t^{\gamma-1}$$

for individuals $i = 1, 2, 3, 4$ where ϕ_i depends in some way on the smoking status and weight.

The next question is: how can we make the constants ϕ_i depend on the covariates. This is done using the following linear approach. We let

$$\begin{aligned} \phi_i &= \exp(\eta_i), \quad \text{where} \\ \eta_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \\ &= \sum_{k=1}^{p} \beta_k x_{ik} \\ &= \underline{\beta}^T \underline{x}_i \end{aligned}$$

for some constants $\beta_1, \ldots, \beta_p$. We call η_i the *linear predictor* for individual i. It is also referred to as the *risk score* or *prognostic index.*

For the study above there are two covariates for each individual:

$$x_{i1} = \begin{cases} 1 \text{ if smoker} \\ 0 \text{ if non-smoker} \end{cases}$$
$$x_{i2} = \text{weight in kg of individual } i.$$

We might decide that $\beta_1 = 0.7$ and $\beta_2 = 0.008$ in which case

$$h_i(t) = h_0(t) \times \exp(0.7 \times x_{i1} + 0.008 \times x_{i2}).$$

Then for example

$$h_1(t) = h_0(t) \times \exp(0.7 \times 1 + 0.008 \times 70).$$

Note that it might be the case that no individual in the study may actually have the same hazard function as the baseline (although this may happen). In the example above the baseline corresponds to the hazard experienced by an individual who is a non-smoker and who has zero weight! The baseline hazard function is only used in *relative terms* in order to obtain a hazard function for any given individual.

Why is this sort of model sensible? First, changes in the explanatory variables *multiply* the hazard. Secondly, the addition of the covariates in the linear expression means that they affect the hazard *independently.*

For example, using the information above, suppose we have 2 individuals who have the same weight w, one of whom is a smoker and the other a non-smoker. Then the hazards are related by:

$$\frac{h_{\text{smoker}}(t)}{h_{\text{non-smoker}}(t)} = \exp(\beta_1) = e^{0.7} \approx 2.$$

In other words the hazard is twice as large for a smoker than a non-smoker at the same weight.

Similarly, we can consider how changes in one covariate affect the hazard: eg. suppose

$$\begin{aligned} \eta_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \\ \eta_2 &= \beta_1 (x_{11} + \delta) + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \end{aligned}$$

so that individual 2 is identical to individual 1 except that x_{11} is changed by an amount δ. Then

$$\frac{h_2(t)}{h_1(t)} = \frac{h_0(t)\exp(\eta_2)}{h_0(t)\exp(\eta_1)} = \exp(\beta_1 \delta).$$

For example if $\beta_2 = 0.008$ and x_{i2} is the weight in kilos of each individual, then the hazard scales by $\exp(0.08) = 1.08$ for every 10kg of weight you put on.

More generally, if we consider individuals i, j with covariates

$$x_{i1}, x_{i2}, \ldots, x_{ip} \quad \text{and} \quad x_{j1}, x_{j2}, \ldots, x_{jp}$$

then

$$\begin{aligned} \frac{h_i(t)}{h_j(t)} &= \frac{h_0(t)\exp(\eta_i)}{h_0(t)\exp(\eta_j)} \\ &= \exp[\beta_1(x_{i1} - x_{j1}) + \beta_2(x_{i2} - x_{j2}) + \cdots + \beta_p(x_{ip} - x_{jp})]. \end{aligned}$$

11.2 Algebraic relationships

Under the proportional hazards assumption the following algebraic relationships hold:

Hazard function:

$$h_i(t) = h_0(t) \times e^{\eta_i}$$

Cumulative hazard:

$$H_i(t) = \int_0^t h_0(t) \exp(\eta_i) dt = H_0(t) \exp(\eta_i)$$

Survivor function:

$$\begin{aligned} S_i(t) &= \exp[-H_i(t)] \\ &= \exp[-H_0(t) e^{\eta_i}] \\ &= [S_0(t)]^{\exp(\eta_i)} \end{aligned}$$

Distribution function:

$$F_i(t) = 1 - [S_0(t)]^{\exp(\eta_i)}$$

Density function:

$$f_i(t) = h_0(t) e^{\eta_i} [S_0(t)]^{\exp(\eta_i)}$$

The survivor functions of two individuals i, j are related in the following way:

$$S_i(t) = [S_0(t)]^{\exp(\eta_i)} \quad \text{and} \quad S_j(t) = [S_0(t)]^{\exp(\eta_j)}$$

so

$$[S_i(t)]^{\exp(\eta_j)} = [S_j(t)]^{\exp(\eta_i)}$$

which gives

$$S_i(t) = [S_j(t)]^{\exp(\eta_i - \eta_j)}$$

11.3 Factors

In the examples above we saw two kinds of covariate:

- Continuous covariates eg. age, weight, BMI
- Indicator variables / logical covariates eg. smoking status, sex

Smoking status and sex are examples of *factors*. These could only take two values but some factors can take more than two values. A factor is a covariate that can adopt a finite number of values which are categorical rather than numerical (even though sometimes they may be given numbers as labels)..

Example: In the project "Operation Type" (`optype`), "Change of Activity" (`ca`) and "Exercise Grade" (`enow`) are factors. "Operation Type" can take three values as can "Change of Activity". Notice that, in the case of "Change of Activity", there is a natural ordering with "No change" coming in between "Decrease" and "Increase". This is not always the case and there is not really any obvious natural ordering of the Operation Types (although, in fact, they were ordered in time). In the case of "Exercise Grade" there are five ordered values. It is typical to represent the level of fitness of an individual by a factor. So, in this case, "Exercise Grade" has 5 different values $0, 1, 2, 3, 4$ where 0 means unfit and 4 means fit.

The different values adopted by a factor are called *levels*. In the exercise grade example, the factor has 5 different levels. Note that an indicator covariate is a factor with 2 levels.

One way to incorporate factors into an analysis is as follows. We pick one level as the baseline (eg. exercise grade 0) and introduce a logical variable for the remaining levels.

Example: Let

$$x_{i1} = \begin{cases} 1 \text{ if individual } i \text{ has exercise grade 1} \\ 0 \text{ if individual } i \text{ does not have exercise grade 1} \end{cases}$$

and define x_{i2}, x_{i3}, x_{i4} similarly to indicate grades 2, 3, and 4 respectively. We then incorporate the covariates $x_{i1}, x_{i2}, x_{i3}, x_{i4}$ into our model as normal, using coefficients $\beta_1, \beta_2, \beta_3, \beta_4$.

An alternative method which pays more attention to the ordering would be to define logical variables as follows, where z_i is the numerical value of the exercise grade for patient i..

$$\begin{aligned} x_{i1} &= \begin{cases} 0 & (z_i = 0) \\ 1 & (z_i \geq 1) \end{cases} \\ x_{i2} &= \begin{cases} 0 & (z_i < 2) \\ 1 & (z_i \geq 2) \end{cases} \\ x_{i3} &= \begin{cases} 0 & (z_i < 3) \\ 1 & (z_i \geq 3) \end{cases} \\ x_{i4} &= \begin{cases} 0 & (z_i < 4) \\ 1 & (z_i = 4) \end{cases} \end{aligned}$$

As you can see from the example, a factor with k levels introduces $k - 1$ coefficients into a model. If you include such a factor in a model in R, the output gives the value of the $k - 1$ coefficients. Here's the output for an example involving "disease stage", treating disease stage as a continuous covariate:

```
> survreg(Surv(my.data$time, my.data$cens)~my.data$stage)

Coefficients:
  (Intercept) my.data$stage
   10.5120199    -0.6786783

Scale= 0.8280779

Loglik(model)= -1162.9   Loglik(intercept only)= -1188.8
        Chisq= 51.7 on 1 degrees of freedom, p= 6.5e-13
```

Here's the output treating disease stage as a factor:

```
> survreg(Surv(my.data$time, my.data$cens)~factor(my.data$stage))

Coefficients:
           (Intercept) factor(my.data$stage)2 factor(my.data$stage)3
             10.291537              -1.302207              -1.750934
factor(my.data$stage)4
             -2.512294

Scale= 0.8272852

Loglik(model)= -1162.3   Loglik(intercept only)= -1188.8
        Chisq= 52.97 on 3 degrees of freedom, p= 1.9e-11
```

While the disease stage *could* be incorporated either as a factor or continuous variate, there is no particularly good reason to assume the change in hazard from stage 1 to 2 is the same as the change in hazard from stage 3 to 4 (for example). Representing disease stage as a factor rather than a continuous covariate allows a model to have different changes in hazard between the disease stages (similarly exercise grades).

12 More on proportional hazards models

In the last section we introduced the theory behind proportional hazards models and how to incorporate covariates with them. The main idea missing was any discussion of how we might go about estimating the coefficients $\beta_1, \ldots, \beta_p$. The next two sections explain how this is done, first for parametric proportional hazards models, and then for the semi-parametric Cox model.

12.1 The Weibull parametric proportional hazards model

In the last section we saw a PH (proportional hazards) model for which the underlying baseline hazard was Weibull:

$$h_i(t) = \phi_i h_0(t)$$

where

$$\begin{aligned} \phi_i &= \exp \eta_i = \exp \underline{\beta}^T \underline{x}_i, \\ h_0(t) &= \lambda \gamma t^{\gamma - 1}. \end{aligned}$$

This is a *parametric* proportional hazards model, since we are assuming a parametric form for the baseline hazard. Note that the hazard function for the i-th individual is another Weibull hazard, but with $\lambda_i = \phi_i \times \lambda$ and $\gamma_i = \gamma$. Obviously, this is no use to us unless we can estimate the parameters in the model: given a set of survival data (with possible right-censoring) we want a way to estimate the coefficients $\underline{\beta}$ as well as the parameters λ and γ.

This can be carried out via maximum likelihood estimation, just like we did in the case with no covariates. In practice, of course, we use R (or another software package) to do this, and we need to understand what R does. Recall that in the case of no covariates we used a transformation of the form

$$\log T = -\frac{1}{\gamma} \log \lambda + \frac{1}{\gamma} \log E$$

for a Weibull life time random variable with parameters λ and γ. Under the Weibull PH model, we have a life time random variable T_i for each individual $i = 1, \ldots, n$ with parameters $\lambda_i = \phi_i \times \lambda$ and $\gamma_i = \gamma$. This gives

$$\log T_i = -\frac{1}{\gamma_i} \log \lambda_i + \frac{1}{\gamma_i} \log E_i$$

where $E_i \sim \exp(1)$. Expanding λ_i and γ_i leads to

$$\begin{aligned}\log T_i &= -\frac{1}{\gamma}\log\lambda - \frac{1}{\gamma}\log\phi_i + \frac{1}{\gamma}\log E_i\\ &= -\frac{1}{\gamma}\log\lambda - \frac{1}{\gamma}(\beta_1 x_{i1} + \cdots + \beta_p x_{ip}) + \frac{1}{\gamma}\log E_i.\end{aligned}$$

The R function `survreg` carries out maximum likelihood estimation via this transformation. It uses the following parameterization:

$$\log T_i = \mu + (\alpha_1 x_{i1} + \cdots + \alpha_p x_{ip}) + \sigma \log E_i$$

where

μ is called the intercept,
σ is called the scale, and
$\alpha_1, \ldots, \alpha_p$ are the coefficients.

To recover the parameters $\underline{\beta}, \lambda, \gamma$ you therefore use

$$\begin{aligned}\gamma &= \sigma^{-1}\\ \lambda &= \exp(-\mu/\sigma)\\ \underline{\beta} &= -\sigma^{-1}\underline{\alpha}.\end{aligned}$$

Example: Fitting a Weibull parametric model. We can fit a model including the patient log bilirubin and presence of ascites as covariates:

```
s <- survreg(formula = Surv(time, cens) ~ ascites+logbilirubin,
data=my.data, dist="weibull")
> summary(s)

Call:
survreg(formula = Surv(time, cens) ~ ascites + logbilirubin,
    data = my.data, dist = "weibull")
              Value Std. Error      z        p
(Intercept)   8.790     0.1056  83.27 0.00e+00
ascites      -0.925     0.1659  -5.57 2.52e-08
logbilirubin -0.652     0.0623 -10.47 1.20e-25
Log(scale)   -0.408     0.0725  -5.62 1.88e-08

Scale= 0.665

Weibull distribution
Loglik(model)= -1110.5   Loglik(intercept only)= -1188.8
        Chisq= 156.58 on 2 degrees of freedom, p= 0
```

12.2 Other parametric models

A key point above was that when we multiply the Weibull hazard function by a constant the result is another Weibull hazard function with a different λ parameter. Consider what happens if we try this with a log-logistic baseline:

$$h_0(t) = \frac{\gamma\rho(\rho t)^{\gamma-1}}{1+(\rho t)^{\gamma}}.$$

If we multiply $h_0(t)$ by a constant the result no longer has the log-logistic form. While you could model data in this way, obviously it isn't really a log-logistic model as only the baseline would have the correct form. In fact, the log-logistic distribution can be used as an *accelerated life* model, as we'll see later.

Note that since the exponential distribution is just Weibull but with $\gamma = 1$, it forms a suitable parametric PH model.

12.3 Comparing nested models

As in a normal multiple linear regression, we can test the contributions of covariates to the model. Specifically, we test the null hypothesis that the coefficients of some subset of the explanatory variables are all zero. This applies to the Cox PH model as well as parametric PH models.

Suppose we fit a model with linear predictor

$$\eta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

We might then add a number of additional covariates $x_{i,p+1}, \ldots, x_{i,p+r}$ to give a second model with

$$\eta_i' = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \beta_{p+1} x_{i,p+1} + \cdots + \beta_{p+r} x_{i,p+r}.$$

Let L and L' be the maximum values of the likelihood for the two models. Then, under the null hypothesis that

$$\beta_{p+1} = \beta_{p+2} = \beta_{p+r} = 0,$$

we have approximately (ie. for large samples)

$$W = (-2\log L) - (-2\log L') \sim \chi_r^2$$

so we compare the statistic W with the upper tail of χ_r^2.

Example: We can compare the Weibull model above with one which does not include log bilirubin:

```
> s <- survreg(formula = Surv(time, cens) ~ ascites,
data=my.data, dist="weibull")
> summary(s)

...

Loglik(model)= -1163.8   Loglik(intercept only)= -1188.8
        Chisq= 49.89 on 1 degrees of freedom, p= 1.6e-12
```

This model has $\log L = -1163.8$, whereas the model with log bilirubin too had $\log L' = -1110.5$. In this case

$$W = 2 \times 1163.8 - 2 \times 1110.5 = 106.6$$

which, compared to χ_1^2, is very significant.

12.4 Confidence intervals for hazard ratios

For large samples, the maximum likelihood estimates of the coefficients are approximately normally distributed. This allows us to construct confidence intervals for hazard ratios.

Example:

Variable	Estimated coefficient	Std. Error
Sex (0: male, 1: female)	$\hat{\beta}_1 = 0.262$	0.017
Age (years)	$\hat{\beta}_2 = 0.023$	0.008

Find a 95% confidence interval for the hazard ratio for a 60-year-old female compared to a 40-year-old female.

Hazard ratio:

$$R = \frac{\exp(\beta_1 + 60\beta_2)}{\exp(\beta_1 + 40\beta_2)} = e^{20\beta_2}$$

Point estimate:

$$\hat{R} = e^{20\hat{\beta}_2} = e^{20\times 0.023} = e^{0.46} = 1.584$$

Confidence interval for β_2:

$$0.023 \pm 1.96 \times 0.008$$

That is

$$0.00732 < \beta_2 < 0.03868$$

Confidence interval for hazard ratio:

$$e^{20\times 0.00732} < e^{20\beta_2} < e^{20\times 0.03868}$$

That is

$$1.158 < R < 2.168$$

Note that this calculation was based on the “β” coefficients, not the “α” coefficients.