QMWS - Survival Analysis
Cox proportional hazards model

Instructor: Kevin McGregor

York University
Department of Mathematics and Statistics

Regression models

We've seen two kinds of regression models for survival data so far:

- Exponential regression
- Weibull regression

Each one assumes the underlying event times follow a particular distribution.

In this lecture, we will discuss the most important type of regression model in survival analysis: the **Cox proportional hazards model**.

Exponential/Weibull regression

Recall the hazard model formulations:

Exponential regression:

$$\lambda_i(t, x_i) = \exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\}$$

Weibull regression:

$$\lambda_i(t, x_i) = \gamma(\lambda t)^{\gamma-1} \exp\left\{\beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right\},$$

Also recall the hazard ratios in each case, where predictor x_{ij} differs by one unit between two individuals, other predictors are constant. $x_1 = (x_{11}, \ldots, a, \ldots, x_{1p})$ and $x_2 = (x_{21}, \ldots, a+1, \ldots, x_{2p})$, $x_{1k} = x_{2k}$ if $k \neq j$.

Exponential regression:

$$\frac{\lambda_2(t, x_2)}{\lambda_1(t, x_1)} =$$

Weibull regression:

$$\frac{\lambda_2(t, x_2)}{\lambda_1(t, x_1)} =$$

Note that the part in front of the $\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$ term cancels out in both cases. This will clearly happen for any hazard model with the following form:

$$\lambda_i(t, x_i) = \lambda_0(t) \exp\{\beta_1 x_{i1} + \cdots + \beta_p x_{ip}\}$$

Hazard ratio:

$$\frac{\lambda_2(t, x_2)}{\lambda_1(t, x_1)} =$$

Cox proportional hazards model

Specifying the following hazard function is leads to the **Cox proportional hazards model**.

$$\lambda_i(t, x_i) = \lambda_0(t) \exp\{\beta_1 x_{i1} + \cdots + \beta_p x_{ip}\}$$

The model is named after **Sir David Cox**, one of the most influential statisticians of all time.

The idea here is that the term $\lambda_0(t)$ remains **unspecified**. We call $\lambda_0(t)$ the **baseline hazard**.

- Baseline hazard is the value of $\lambda_i(t, x_i)$ when all covariates in x_i are equal to 0.
- $\lambda_0(t)$ is called a **nuisance parameter**.

Cox proportional hazards model

$$\lambda_i(t, x_i) = \lambda_0(t) \exp\{\beta_1 x_{i1} + \cdots + \beta_p x_{ip}\}$$

Two important things:

- The hazard depends on time only through the baseline hazard $\lambda_0(t)$.
- The hazard depends on the predictors only through $\exp\{\beta_1 x_{i1} + \cdots + \beta_p x_{ip}\}$.
- **There is no intercept term. E.g. no β_0**

The predictors are assumed to have a **multiplicative** effect on the hazard. This part of the model is **parametric**. The baseline hazard $\lambda_0(t)$ is unspecified, and therefore, **non-parametric**.

- Combining these two terms means this model is **semi-parametric**.

Advantages/disadvantages

Advantages of the proportional hazards model:

- Don't have to specify the baseline hazard.
- Straightforward interpretation of parameters.
- Hazard function can differ with time, but we don't have to worry that part, since only the baseline hazard depends on time.

Disadvantages of the proportional hazards model:

- Since $S(t)$ depends on $\lambda_0(t)$, which is unspecified, then we can't directly estimate survival.
  - Usually when we run the Cox model, we're interested in how the hazard changes with respect to covariates, not the survival function itself.
- The effect of the predictors x_i can't depend on time. Luckily this can be changed, though it leads to a different kind of model. More on this later.

Parameter interpretations

How to interpret each slope parameter β_j? Same as before:

When predictor j is increased by **one unit**, the hazard is multiplied by a factor of $\exp\{\beta_j\}$. This value is called the **hazard ratio**.

- Hazard ratio > 1 implies increasing predictor corresponds to increasing hazard.
- Hazard ratio < 1 implies increasing predictor corresponds to decreasing hazard.

Remember, there is no intercept (β_0) in the Cox model.

The baseline hazard $\lambda_0(t)$ is the hazard at time t when all predictors are equal to 0.

Proportional hazards assumption

What does **proportional hazards** mean?

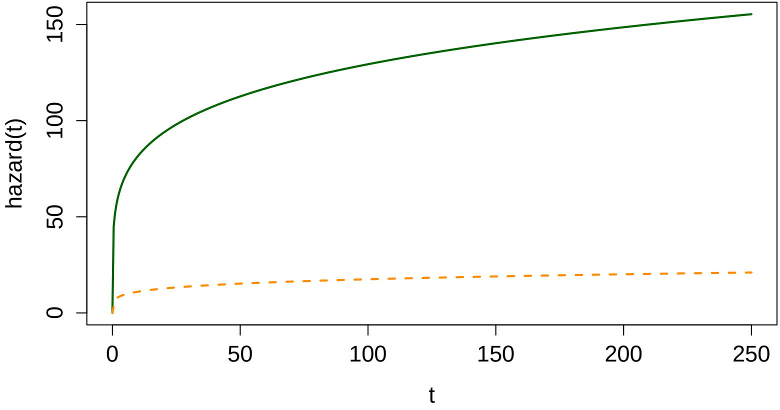
Hazard ratio between two individuals with different predictors:

$$\frac{\lambda_2(t, x_2)}{\lambda_1(t, x_1)} = \frac{\exp\{\beta_1 x_{21} + \cdots + \beta_p x_{2p}\}}{\exp\{\beta_1 x_{11} + \cdots + \beta_p x_{1p}\}}$$

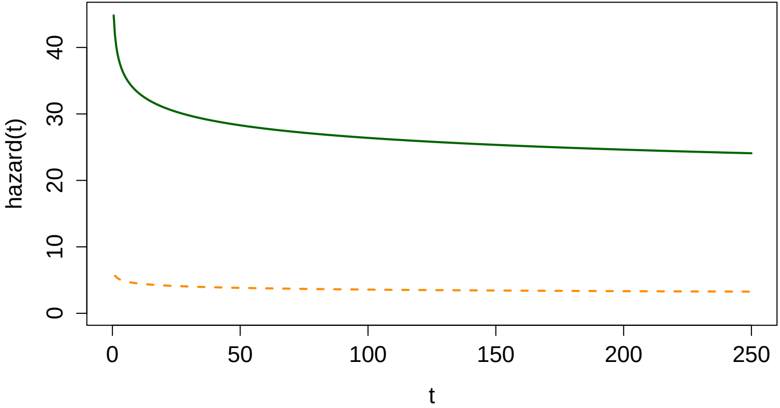
This does not depend on time.

By using this model, we are assuming that the ratio of hazards for two individuals with different predictor profiles remain proportional over time.

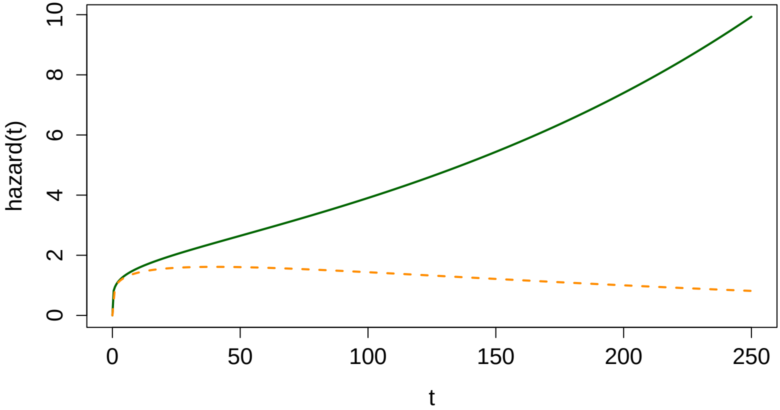
Proportional hazard example 1



Proportional hazard example 2



NOT proportional hazard example



Veterans example

Example: Veteran lung cancer dataset. Outcome is death. Two covariates we'll use in the model are:

- `age`: Ranges from 34-81 years.
- `karno`: Karnofsky score. Higher value corresponds to higher ability for a patient to care for themselves. Very low score refers to hospitalization.

We'll fit a Cox proportional hazards model with `age` and `karno` as predictors. (We'll see how to do this in R)

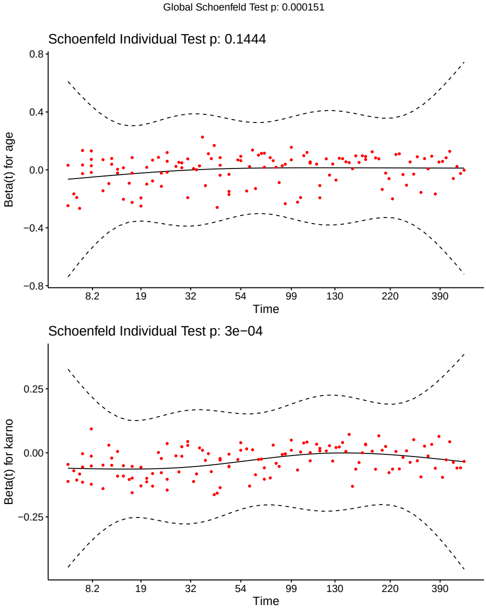
Veterans example

```
coxph(formula = Surv(time, status) ~ age + karno, data = veteran)

  n= 137, number of events= 128

           coef exp(coef)  se(coef)      z Pr(>|z|)
age   -0.002392  0.997611  0.009077 -0.263    0.792
karno -0.033707  0.966855  0.005199 -6.484 8.94e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hypothesis test: proportional hazards assumption

	chisq	df	p
age	2.13	1	0.14439
karno	13.07	1	0.00030
GLOBAL	17.60	2	0.00015

Testing proportional hazards assumption



Proportional hazards discussion

A few points about proportional hazards models:

- We don't specify the baseline hazard. If we want to estimate the survival function itself $S(t)$, then we need to specify a particular model for the baseline hazard $\lambda_0(t)$.
- We make the **proportional hazards** assumption. This is a fairly strong assumption.
- If we have a predictor that we don't think fulfills the proportional hazards assumption, we can do **stratification**.
- We can modify the model so that the predictor effects depend on time.
    - This is called an **accelerated failure-time** model. Not covered in this workshop.

Stratification

In the Cox model, we have the **proportional hazards** assumption. What do we do if this assumption is not met for a particular predictor?

One way to deal with this is by **stratification**. This allows us to assume separate baseline hazards for different values of this predictor.

- We account for the effects of this predictor on the hazard, but there's no β parameter corresponding to that predictor.

Note: stratification can only be done for a **categorical** predictor variable.

Stratification

Suppose we have a predictor x_{ij}, where j denotes a categorical variable with J levels:

$$x_{ij} \in \{1, 2, \ldots, J\}$$

This is called the **stratum** for individual i. The strata could represent different data centres, different hospitals, gender, etc.

We then assume that the following hazard for individual i who is in stratum j:

$$\lambda_{ij}(t) = \lambda_{0j}(t) \exp\{\beta_1 x_{i1} + \cdots + \beta_p x_{ip}\}$$

That is, we assume a distinct baseline hazard $\lambda_{0j}(t)$ in each stratum j.

Stratification: discussion

Advantage of stratification:

- Accounts for a variable that does not meet the proportional hazards assumption
- Relatively straightforward implementation.

Disadvantages of stratification:

- Can't estimate the effect of the stratified variable.
  - We only use stratification if we don't care about that effect.
- Over-stratification can lead to loss of efficiency of estimation for β.
- Cannot be used with a continuous predictor, unless it is first categorized.

Cox proportional hazards: likelihood

Reviewing notation in the Cox proportional hazards model:

We have n event times:

$$t_1, t_2, \ldots, t_n$$

Before, we assumed no ties. We can actually relax this assumption and assume that there are d_i events at each time t_i, for $i = 1, \ldots, n$.

Second, at each time point t_i, define the **risk set** to be the set of individuals in the study that have not yet had an event and have not yet been censored. Denote the risk set at time t_i by:

$$\mathcal{R}_i = \text{Risk set at time } t_i$$

Breslow estimator for cumulative hazard

In the basic Cox proportional hazards model, we can't estimate the survival curve, since we don't specify the baseline hazard $\lambda_0(t)$.

- To estimate survival, we first have to **estimate** the baseline hazard.

One possible estimator for baseline hazard:

$$d\widehat{\Lambda}_0(t_i) = \frac{d_i}{\sum_{j \in \mathcal{R}_i} \exp\{\beta_1 x_{j1} + \cdots + \beta_p x_{jp}\}}$$

This estimator is set to 0 between event times.

Breslow estimator for cumulative hazard

Estimator for cumulative baseline hazard:

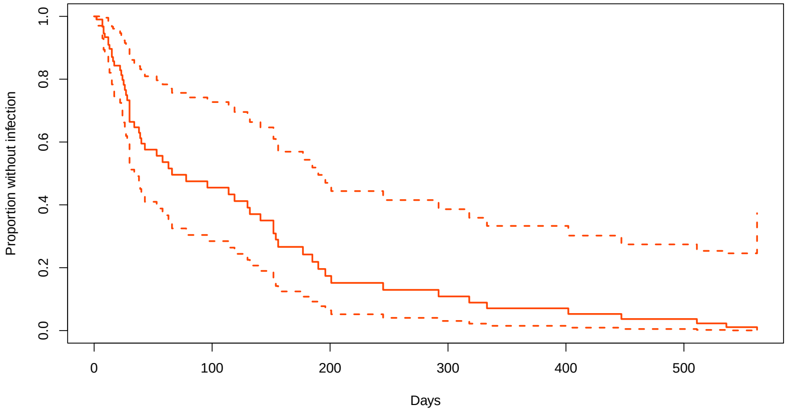
$$\widehat{\Lambda}_0(t) = \sum_{i:t_i \leq t} \frac{d_i}{\sum_{j \in \mathcal{R}_i} \exp\{\beta_1 x_{j1} + \cdots + \beta_p x_{jp}\}}$$

This is called the **Breslow estimator**.

We can then estimate the survival function from the model as:

$$\begin{aligned}\widehat{S}_i(t) &= \exp\left\{-\widehat{\Lambda}_i(t)\right\} \\ &= \exp\left\{-\widehat{\Lambda}_0(t) \exp\{\beta_1 x_{j1} + \cdots + \beta_p x_{jp}\}\right\}\end{aligned}$$

Breslow estimator: time to kidney infection



Log-log plots

A common way to check the proportional hazards assumption is using a **log-log** plot.

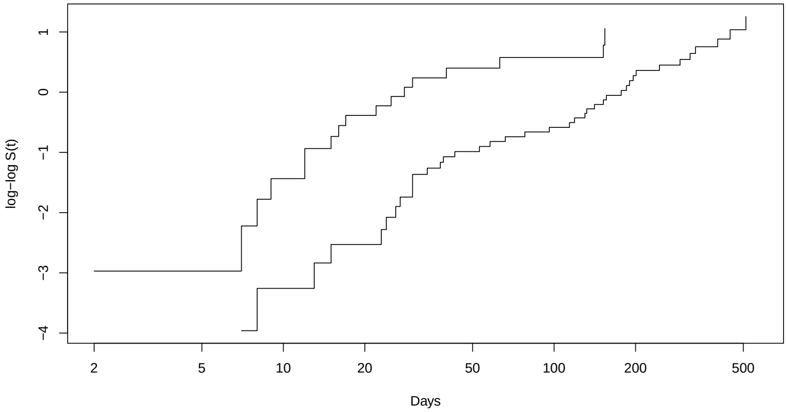
Consider two individuals with predictor profiles x_1 and x_2. Their survival functions are $S_1(t, x_1)$ and $S_2(t, x_2)$, respectively. Look at the following expressions:

$$-\log\left(-\log\left[S_1(t, x_1)\right]\right) \qquad \text{and} \qquad -\log\left(-\log\left[S_2(t, x_2)\right]\right)$$

The **difference** between the transformed curves should not change with time. Thus, we look to see if the curves are **parallel**.

- To keep it simple, we usually just use Kaplan-Meier estimates of survival to create the log-log plot.

Log-log plot: time to kidney infection. Two curves correspond to male/female.



Log-log plot

Log-log plot: time to kidney infection. Two curves correspond to age greater than or less than median age.