Survival Regression

So far we have studied a single set of survival data, or comparing multiple groups of survival times. Often, there are other explanatory variables (or predictors) recorded in a study.
Example: Overall survival of myeloma patients

- Outcome: survival times of 48 patients
- Predictors: age, sex, blood urea nitrogen, serum calcium, serum hemoglobin, percentage of plasma cells, Bence-Jones protein.

In order to evaluate the association between predictors and survival, some kind of regression modeling is needed.

Two types of modeling strategies are frequently used to describe the association between covariates (independent variables) and a failure time variable (dependent variable):

- Parametric approach (e.g., accelerated failure time model)
- Semi-parametric approach (e.g., proportional hazards model)

Accelerated Failure Time (AFT) Model

Let T_i denote the time-to-event variable for the ith subject, and X_i denote a vector of covariates. The accelerated failure time model is

$$\log(T_i) = X_i\beta + \epsilon_i$$

where β is a vector of coefficients and ϵ_i is the random disturbance with some parametric distribution.

- log is the most commonly used transformation, but others are possible
- ϵ_i characterizes the distribution of some underlying unmoderated lifetime (i.e., $\log(T_{0i})$).

The AFT model directly models the effect of explanatory variables on the survival time, so the interpretation of the coefficients are straightforward.

- $T_i = \exp(X_i\beta)T_{0i}$
- Popular distributions for T_{0i} include Weibull, Exponential, and Log-normal.
- Each coefficient β_j indicates for a given individual, with a unit increase in x_{ij}, the expected time-to-event is $\exp(\beta_j)$ times as fast as the original.
- R function survreg can be used to fit parametric models.

Proportional Hazards (PH) Model

Consider a simple example where $X = 1$ for treated subjects and $X = 0$ for controls.

Hazard functions for the two groups:

$$\text{treated: } h_1(t)$$

$$\text{control: } h_0(t)$$

PH model assumes

$$h_1(t) = h_0(t)e^{\beta}$$

This implies that the ratio of the two hazards is a constant, which does NOT depend on t:

$$\frac{h_1(t)}{h_0(t)} = e^{\beta}$$

β is interpreted as the log hazard ratio.

Proportional Hazards (PH) Model

Assume covariates are available on each individual. Let X_i denote the $p \times 1$ vector of covariates for subject i

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T.$$

The PH model assumes that, conditioning on the observed covariates, the hazard function is given by

$$\begin{aligned} h_i(t) &= h_0(t) e^{\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip}} \\ &= h_0(t) e^{X_i^T \beta}, \end{aligned}$$

where β is a $p \times 1$ vector of coefficients.

Interpretation of the model:

Hazard at t for given X_i = (baseline hazard at t) $\times$ (Risk factor $e^{X_i \beta}$)

Features of PH

- PH is based on the hazard function, in contrast with the AFT model based on survival time.
- $h_0(t)$ is the baseline hazard function, which is unspecified (nonparametric).
- The log *hazard ratio* for a subject X to the reference group is $X\beta$ (parametric).
- Overall, PH is semiparametric, with $h_0(t)$ and β to be estimated.
- In most cases, β is of primary interest.
- No intercept in $X\beta$.

Example

T: time from HIV infection to AIDS

- x_{i1}: gender, 0 for female and 1 for male
- x_{i2}: CD4 count measured at baseline for subject i

Consider the univariate PH model: $h_i(t) = h_0(t)e^{\beta x_{i1}}$

Hazard for a female subject at time t: $h_0(t)e^{\beta \cdot 0} = h_0(t)$

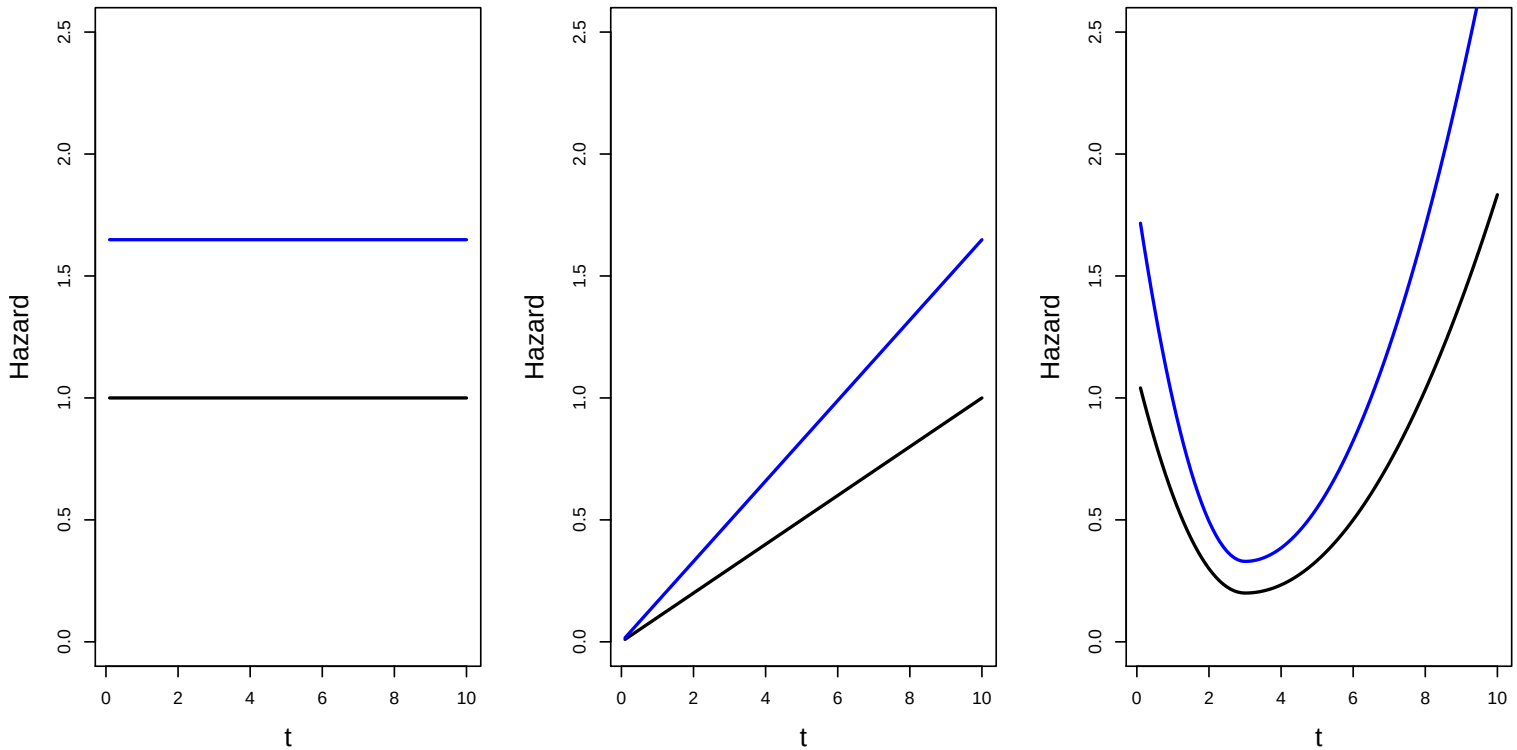
Hazard for a male subject at time t: $h_0(t)e^{\beta \cdot 1} = h_0(t)e^{\beta}$

Relative hazard (hazard ratio) between male and female is

$$\frac{h_0(t)e^{\beta}}{h_0(t)} = e^{\beta}$$

(constant over time)

PH model with different baseline hazard functions $h_0(t)$ and fixed log hazard ratio $\beta = 0.5$ (or HR $= e^{0.5} = 1.65$)



Consider the multivariate PH model:

$$h_i(t) = h_0(t)e^{\beta_1 x_{i1} + \beta_2 x_{i2}}$$

Relative hazard (hazard ratio) for subjects i and k at time t:

$$\begin{aligned} \text{HR}(t) &= \frac{h_i(t)}{h_k(t)} = \frac{h_0(t)e^{x_i\beta}}{h_0(t)e^{x_k\beta}} = e^{(x_i - x_k)\beta} \\ &= e^{\beta_1(x_{i1} - x_{k1}) + \beta_2(x_{i2} - x_{k2})} \end{aligned}$$

With $x_{i1} = x_{k1}$ (same gender), if $\beta_2 = -0.01$, $x_{i2} = 250$, $x_{k2} = 200$, then

$$\text{HR}(t) = e^{-0.01 \times (250-200)} = e^{-0.5} \approx 0.61.$$

Thus, conditioning on the same gender and survival beyond t, the probability that subject i is diagnosed with AIDS at t is 61% of the probability that subject k is diagnosed with AIDS at t.