Biostatistics (MATH11230)

Introduction to survival analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2022/2023

Proportional hazards model

General context

- We end the course discussing regression models for time to event/survival data.
- There are various ways for modelling the dependency between the event time and the factors that might affect it.
- The two most common approaches are known as the **proportional hazards model** and the **accelerated failure time model**.
- Due to time constraints, here we will only cover the first class of methods: proportional hazards models, undeniably the most popular and widely used regression approach for survival data.

Proportional hazards model

Comparing two groups

- Let us start with the simplest example: suppose that two groups of (male) patients receive either a a standard treatment or a new treatment for prostate cancer.
- The primary goal is to identify whether patients in the two treatment groups have a different survival experience.
- Let $h_S(t)$ and $h_N(t)$ be the hazards of death at time t for patients in the standard and new treatment groups, respectively.
- A convenient model is to assume that the hazard at time t for a patient in the new treatment group is proportional to the hazard, at the same time t, in the standard treatment.
- This proportional hazards model can be expressed as

$$h_N(t) = \psi h_S(t), \qquad t \geq 0, \tag{1}$$

where ψ is a constant.

Proportional hazards model

Comparing two groups

- An implication of the proportional hazards model is that the corresponding true survival functions for individuals in the new and in the standard treatments do not cross.

- To see why, recall that by assumption $h_N(t) = \psi h_S(t)$. Integrating both sides of this expression, multiplying by (-1) and exponentiating gives

$$\exp\left\{-\int_0^t h_N(u)\mathrm{d}u\right\} = \exp\left\{-\int_0^t \psi h_S(u)\mathrm{d}u\right\}.$$

- Further note that

$$\exp\left\{-\int_0^t \psi h_S(u)\mathrm{d}u\right\} = \left[\exp\left\{-\int_0^t h_S(u)\mathrm{d}u\right\}\right]^{\psi},$$

and remember that

$$S(t) = \exp\{-H(t)\} = \exp\left\{-\int_0^t h(u)\mathrm{d}u\right\}.$$

Proportional hazards model

Comparing two groups

- Therefore, if $S_N(t)$ and $S_S(t)$ are the survival functions for the two groups, then

$$S_N(t) = [S_S(t)]^{\psi} .$$

- The parameter ψ is called the **hazard ratio** or **relative hazard**: it is the ratio of the hazard of death at any time for an individual on the new treatment relative to an individual on the standard treatment.

- If $0 < \psi < 1$, the hazard of death at time t is smaller for an individual on the new drug relative to an individual on the standard treatment.

- The new treatment is then an improvement on the standard treatment.

- On the other hand, if $\psi > 1$, the hazard of death at time t is greater for an individual in the new drug, and the standard treatment is superior.

- A very nice short video explaining hazard ratios by Professor Sir David Spiegelhalter

`www.youtube.com/watch?v=BHsdpg1OOf0`

Proportional hazards model

Comparing two groups

- An alternative way of expressing the model in (1) leads to a model that can be more easily generalised.
- Suppose that survival times are available on n individuals and denote the hazard function for the ith of these by $h_i(t)$, for $i = 1, \ldots, n$.
- Also, let $h_0(t)$ be the hazard function for an individual in the standard treatment group.
- The hazard function for an individual on the new treatment is then $\psi h_0(t)$.
- The relative hazard cannot be negative and so it is convenient to set $\psi = \exp(\beta)$.
- The parameter β is then logarithm of the hazard ratio and any value of $\beta \in (-\infty, +\infty)$ will lead to a positive value of ψ.

Proportional hazards model

Comparing two groups

- Now let X be an indicator variable, which takes the value zero if an individual is on the standard treatment and one if the individual is on the new drug.

- The hazard function for individual i is then written as

$$h_i(t) = h(t \mid X = x_i) = h_0(t)e^{\beta x_i}$$
$$= \begin{cases} h_0(t), & \text{if } x_i = 0 \\ h_0(t)e^{\beta}, & \text{if } x_i = 1 \end{cases}$$

- This is the proportional hazards model for the comparison of two treatment groups.

- Here $\psi = e^{\beta}$ represents the ratio of the hazard of death for an individual on the new treatment relative to one on the standard treatment at any time t.

Proportional hazards model

General model

- The model we have considered just now can be generalised to the situation where the hazard of death (or, more generally, the hazard of the event of interest occurring) at a particular time point depends on the values of $x_1, \ldots, x_p$ of p explanatory variables/risk factors/covariates $X_1, \ldots, X_p$.

- The values of these variables will be assumed to have been recorded at the time origin of the study (or in epidemiologic jargon: *at baseline*).

- We will not cover them but there are extensions of the model to cover the situation where the values of one or more explanatory variables change over time.

- Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ be the explanatory variables for individual i.

Proportional hazards model

General model

- The proportional hazards model is given by

$$h_i(t) = h(t \mid \mathbf{x}_i) = h_0(t) \exp(x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p) = h_0(t) \exp(\mathbf{x}_i'\boldsymbol{\beta}), \qquad (2)$$

  where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$.

- Note that there is no intercept β_0 in model (2).

- Obviously,

$$h(t \mid \mathbf{x}_i = \mathbf{0}) = h_0(t),$$

  and, therefore, $h_0(t)$ is often called the baseline hazard function.

- It can be interpreted as the hazard function for a standard subject, which is a subject with $\mathbf{x} = \mathbf{0}$.

Proportional hazards model

General model

- ↪ Any parametric hazard function can be used for $h_0(t)$ and as we will see in the next slides, $h_0(t)$ can be left completely unspecified without sacrificing the ability to estimate β, by the use of Cox's semiparametric proportional hazards model.

- ↪ The hazard function for the ith subject always has the same general shape as $h_0(t)$, but can be, say, doubled or halved, depending on the individual's risk factors.

- ↪ The term $\exp(\mathbf{x}_i'\beta)$ is in many cases the quantity of primary interest as it describes the (relative) effects of the risk factors.

- ↪ Note that the model separates clearly the effect of time from the effect of the risk factors.

Proportional hazards model

General model

↪ The proportional hazards model can also be written in terms of the survival function:

$$\begin{aligned} S_i(t) = S(t \mid \mathbf{x}_i) &= \exp\left(-\int_0^t h_i(u)\mathrm{d}u\right) \\ &= \exp\left(-\int_0^t h_0(u)\exp(\mathbf{x}_i'\boldsymbol{\beta})\mathrm{d}u\right) \\ &= \left[\exp\left(-\int_0^t h_0(u)\mathrm{d}u\right)\right]^{\exp(\mathbf{x}_i'\boldsymbol{\beta})} \\ &= S_0(t)^{\exp(\mathbf{x}_i'\boldsymbol{\beta})}, \end{aligned}$$

where $S_0(t)$ is the baseline survival function.

Proportional hazards model

General model

- On the log-hazard scale this model can be written as

$$\log h_i(t) = \log h_0(t) + \mathbf{x}_i'\boldsymbol{\beta}.$$

- The coefficient β_j associated with the risk factor X_j represents the change in the log hazard when X_j is increased by one unit, all the other risk factors being constant, that is,

$$\begin{aligned}\beta_j = &\log h(t \mid X_1 = x_1, X_2 = x_2, \ldots, X_j + 1 = x_j + 1, \ldots, X_p = x_p) \\ &- \log h(t \mid X_1 = x_1, X_2 = x_2, \ldots, X_j = x_j, \ldots, X_p = x_p),\end{aligned}$$

  which is equivalent to the log of the ratio of the hazards at time t.

- Equivalently,

$$\frac{h(t \mid x_1, x_2, \ldots, x_j + 1, \ldots, x_p)}{h(t \mid x_1, x_2, \ldots, x_j, \ldots, x_p)} = e^{\beta_j}.$$

- Therefore, e^{β_j} represents the ratio of the hazards for a one unit change in X_j at any time t while keeping all the other risk factors constant.

Proportional hazards model

General model

- The main assumption of this model is the **proportional hazards assumption**, that is, the fact that the ratio of the hazard functions for two subjects with covariates $\mathbf{x}_i$ and $\mathbf{x}_l$ is constant over time, that is,

$$\frac{h(t \mid \mathbf{x}_i)}{h(t \mid \mathbf{x}_l)} = \frac{\exp(x_i'\beta)}{\exp(x_l'\beta)} = \exp\{(\mathbf{x}_i - \mathbf{x}_l)'\beta\}.$$

- In other words, all the time dependency is captured by the baseline hazard function which is common to all observations.

- The proportional hazards model also further assumes that the relationship between the risk factors and the log hazard function is linear (for the case of continuous covariates). This can be relaxed but it is out of the scope of the course.

Proportional hazards model

Toy example (from Klein and Moeschberger, 2003, p248)

- Let us consider for the sake of simplicity/illustration that the only risk factor that needs to be considered is race and that we are interested in the time until death.
- Let us consider that race is a three level factor, whose levels are: black, white, hispanic.
- Define

$$X_1 = \begin{cases} 1, & \text{if the subject is black,} \\ 0, & \text{otherwise,} \end{cases} \qquad X_2 = \begin{cases} 1, & \text{if the subject is white,} \\ 0, & \text{otherwise.} \end{cases}$$

- Hispanic is therefore the reference category.

Proportional hazards model

Toy example (from Klein and Moeschberger, 2003, p248)

- The proportional hazards model is

$$h(t \mid X_1 = x_1, X_2 = x_2) = h_0(t) \exp(x_1\beta_1 + x_2\beta_2)$$
$$= \begin{cases} h_0(t), & \text{if } X_1 = X_2 = 0 \text{ (hispanic subject)}, \\ h_0(t)\exp(\beta_1), & \text{if } X_1 = 1, X_2 = 0 \text{ (black subject)}, \\ h_0(t)\exp(\beta_2), & \text{if } X_1 = 0, X_2 = 1 \text{ (white subject)}. \end{cases}$$

- The hazard of death among black subjects relative to the hazard of death occurring among hispanic subjects is e^{β_1}.

- The hazard of death among white subjects relative to the hazard of death among hispanic subjects is e^{β_2}.

- The hazard of death among black subjects relative to the hazard of death among white subjects is $e^{\beta_1 - \beta_2}$.

Proportional hazards model

Fully parametric specification

- Choosing a parametric distribution for the event times imposes a parametric form on $h_0(t)$ and leads to a fully parametric proportional hazards model.

- Remember that under a parametric setup, we have seen last week that the likelihood of the parameters of the model, say θ, is given by

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} h(y_i;\theta)^{\delta_i} S(y_i;\theta) \\ &= \prod_{i=1}^{n} [h_0(y_i;\theta_1)\exp(\mathbf{x}_i'\beta)]^{\delta_i} S_0(y_i;\theta_1)^{\exp(\mathbf{x}_i'\beta)}, \quad \theta = (\theta_1, \beta). \end{aligned}$$

  where $y_i = \min\{t_i, c_i\}$ and δ_i is the censoring indicator.

- Once a given distribution has been chosen for T, one can plug in the corresponding baseline hazard and survival functions in the above likelihood.

Proportional hazards model

Fully parametric specification

- For instance, considering that the survival times follow a Weibull distribution with scale parameter λ and shape parameter α, then the baseline hazard and survival functions are given by

$$h_0(t) = \lambda\alpha t^{\alpha-1}, \qquad S_0(t) = e^{-\lambda t^{\alpha}}, \qquad \lambda > 0, \alpha > 0.$$

- The corresponding hazard and survival functions for the ith individual with covariates $\mathbf{x}_i$ are given by

$$h(t \mid \mathbf{x}_i) = h_0(t)\exp(\mathbf{x}_i'\boldsymbol{\beta}) = \lambda\alpha t^{\alpha-1}\exp(\mathbf{x}_i'\boldsymbol{\beta})$$
$$S(t \mid \mathbf{x}_i) = S_0(t)^{\exp(\mathbf{x}_i'\boldsymbol{\beta})} = e^{-\lambda t^{\alpha}\exp(\mathbf{x}_i'\boldsymbol{\beta})}$$

- Letting $\theta = (\lambda, \alpha, \boldsymbol{\beta})$, $y_i = \min\{t_i, c_i\}$, and δ_i to be the censoring indicator, the likelihood is then written by

$$L(\theta) = \prod_{i=1}^{n} \left\{ \left[\lambda\alpha y_i^{\alpha-1}\exp(\mathbf{x}_i'\boldsymbol{\beta})\right]^{\delta_i} e^{-\lambda y_i^{\alpha}\exp(\mathbf{x}_i'\boldsymbol{\beta})} \right\},$$

with the corresponding log likelihood being given by

$$\log L(\theta) = \sum_{i=1}^{n} \left\{ \delta_i \left[\log\lambda + \log\alpha + (\alpha-1)\log y_i + \mathbf{x}_i'\boldsymbol{\beta}\right] - \lambda y_i^{\alpha}\exp(\mathbf{x}_i'\boldsymbol{\beta}) \right\}.$$

Proportional hazards model

Fully parametric specification

- ↪ In order to obtain the maximum likelihood estimate of θ we need to maximise the log likelihood in the previous slide, which has to be done numerically.

- ↪ The maximisation of the log likelihood, regardless of the parametric distribution chosen, is almost always, if not always, performed numerically.

Proportional hazards model

Semiparametric specification

- In the early seventies, Sir David Cox proposed a proportional hazards model in which the baseline hazard function $h_0(t)$ was left totally unspecified and showed that we can still obtain estimators for the regression coefficients $\boldsymbol{\beta}$.

- When the baseline hazard function is left unspecified, the likelihood function in (1) can no longer be directly maximised to find the estimates of $\boldsymbol{\beta}$.

- The idea is then to maximise a **partial likelihood** with respect to $\boldsymbol{\beta}$ that no longer involves $h_0(t)$.

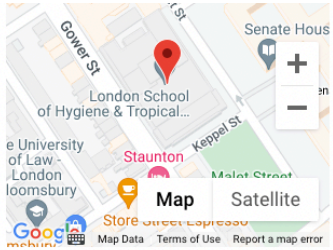
- This model specification is referred to as **semiparametric**. The parametric part of the model is the term $\exp(\mathbf{x}'\boldsymbol{\beta})$, but we make no assumptions about $h_0(t)$.

Proportional hazards model

Semiparametric specification

A celebration of 50 Years of the Cox model in memory of Sir David Cox

Share



Venue LSHTM, Keppel Street
London
WC1E 7HT
United Kingdom

Get Directions

Room John Snow Lecture Theatre and South Courtyard Café

Date Thursday 10 November 2022

Time 11:00 - 19:30

Date and time zone is UK

Admission

Proportional hazards model

Semiparametric specification

- Let us first assume the case in which there are no ties in the data.
- Let $t_1 < t_2 < \ldots < t_J$ be the ordered event times.
- The set of individuals who are at risk at time t_j is denoted by $R(t_j)$ and is known as the risk set.
- Cox (1972) showed that the partial likelihood is given by

$$Lp(\beta) = \prod_{j=1}^{J} \frac{\exp(\mathbf{x}_j'\beta)}{\sum_{l \in R(t_j)} \exp(\mathbf{x}_l'\beta)} \tag{3}$$

where $\mathbf{x}_j$ is the vector of covariates for the individual that experiences the event at time t_j.

Proportional hazards model

Semiparametric specification

- ↪ Note that the sum in the denominator is over all individuals who are at risk at time t_j and the contribution of censored observations is only through the denominator (they do not contribute to the numerator) in the risk sets of event times that occur before a censored time.

- ↪ We can also notice that this partial likelihood only depends on the order of the events, since this determines the risk set at each event time, and not on the timing of the events themselves.

Proportional hazards model

Semiparametric specification

- ↪ The basis of the argument used in the construction of this partial likelihood is that intervals between successive event times convey no information about the effect of the explanatory variables/risk factors on the hazard of death.
- ↪ This is because the baseline hazard function has an arbitrary form, and so it is conceivable that $h_0(t)$, and hence $h(t \mid \mathbf{x}_j)$, is zero in those time intervals in which there are no events.
- ↪ This in turn means that these intervals give no information about the parameters $\boldsymbol{\beta}$.
- ↪ In other words, we would expect the ranks to contain most of the information about $\boldsymbol{\beta}$ and that, say, the duration of time between t_j and t_{j+1} probably would not add a great deal of information to our knowledge of $\boldsymbol{\beta}$.
- ↪ Anyhow, drawing further conclusions about $\boldsymbol{\beta}$ based on the gaps between failure times is going to be highly dependent on making distributional assumptions concerning h_0.
- ↪ Focusing on the ranks, therefore, is likely to be both efficient and robust.

Proportional hazards model

Semiparametric specification

↪ We therefore consider the probability that the ith individual dies at some time t_j, conditional on t_j being one of the observed set of J death times, $t_1, t_2, \ldots, t_J$. If the vector of values of the risk factors for the individual who dies at t_j is denoted by $\mathbf{x}_j$, then this probability is

$$\begin{aligned}&\Pr(\text{individual with risk factors } \mathbf{x}_j \text{ dies at } t_j \mid \text{one death at } t_j)\\ =&\frac{\Pr(\text{individual with risk factors } \mathbf{x}_j \text{ dies at } t_j)}{\Pr(\text{one death at } t_j)}.\end{aligned} \tag{4}$$

↪ Remember that we are assuming that there are no ties between the uncensored event times $t_1, t_2, \ldots, t_J$.

↪ Because the death times are assumed to be independent of one another, the denominator is the sum of the probabilities of death at time t_j over all individuals who are at risk of death at that time. If these individuals are indexed by l, with $R(t_j)$ denoting the set of individuals who are at risk at time t_j, Expression (4) becomes

$$\frac{\Pr(\text{individual with risk factors } \mathbf{x}_j \text{ dies at } t_j)}{\sum_{l \in R(t_j)} \Pr(\text{individual } l \text{ dies at } t_j)}. \tag{5}$$

Proportional hazards model

Semiparametric specification

- The probabilities of death at time t_j in Expression (5) are now replaced by probabilities of death in the interval $(t_j, t_j + \Delta t)$ and dividing both the numerator and denominator of Expression (5) by Δt, we get

$$\frac{\Pr(\text{individual with risk factors } \mathbf{x}_j \text{ dies in } (t_j, t_j + \Delta t))/\Delta t}{\sum_{l \in R(t_j)} \Pr(\text{individual } l \text{ dies in } (t_j, t_j + \Delta t))/\Delta t}.$$

- The limiting value of this expression as $\Delta t \to 0$ is then the ratio of the probabilities in Expression (5). But, this limit is also the ratio of the corresponding hazards of death at time t_j, that is,

$$\begin{aligned}
&\frac{\text{Hazard of death at time } t_j \text{ for individual with risk factors } \mathbf{x}_j}{\sum_{l \in R(t_j)} \text{Hazard of death at time } t_j \text{ for individual l}} \\
&= \frac{h(t_j \mid \mathbf{x}_j)}{\sum_{l \in R(t_j)} h(t_j \mid \mathbf{x}_l)} \\
&= \frac{h_0(t_j) \exp(\mathbf{x}_j' \boldsymbol{\beta})}{\sum_{l \in R(t_j)} h_0(t_j) \exp(\mathbf{x}_l' \boldsymbol{\beta})} \\
&= \frac{\exp(\mathbf{x}_j' \boldsymbol{\beta})}{\sum_{l \in R(t_j)} \exp(\mathbf{x}_l' \boldsymbol{\beta})}
\end{aligned}$$

Proportional hazards model

Semiparametric specification

- Finally, taking the product of these conditional probabilities over the J (uncensored) death times gives the partial likelihood function in Equation (3).

- As an alternative, it is often convenient to express the likelihood as a product of terms for each time on study (as opposed to only the event/uncensored times). Supposing that there are n times (censored and uncensored), and using the notation from slide 2, we have

$$L_p(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\mathbf{x}_i'\beta)}{\sum_{l \in R(y_i)} \exp(\mathbf{x}_l'\beta)} \right\}^{\delta_i} \tag{6}$$

$$= \prod_{i=1}^{n} \left\{ \frac{\omega_i}{\sum_{l \in R(y_i)} \omega_l} \right\}^{\delta_i}, \quad \omega_i = \exp(\mathbf{x}_i'\beta).$$

- Expressing the partial likelihood as in the last line in the above expression emphasises the fact that the model assigns weight ω_i to the likelihood that individual i will have the event relative to the other subjects at risk.

- Note that the δ_i exponent ensures that only the observations at which an event is observed contribute to the likelihood.

Proportional hazards model

Semiparametric specification

- Note that this is not exactly a likelihood as it does not specify the probability of observing $\{(y_i, \delta_i)\}_{i=1}^n$ given $\{\mathbf{x}_i\}_{i=1}^n$ and $\boldsymbol{\beta}$.

- However, Sir David Cox demonstrated we can maximise this partial likelihood in order to obtain estimates of $\boldsymbol{\beta}$, say $\widehat{\boldsymbol{\beta}}$ and that these estimates are consistent and asymptotically distributed with mean $\boldsymbol{\beta}$ and variance given by the inverse of the observed Fisher information matrix.

- Based on the asymptotic normality (of the vector of) the maximum likelihood estimates, one can also easily construct a $100(1-\alpha)\%$ confidence interval for a particular component of $\boldsymbol{\beta}$, say β_j

$$\widehat{\beta}_j \pm z_{\alpha/2}\sqrt{\widehat{\text{var}}(\widehat{\beta}_j)},$$

where $\widehat{\text{var}}(\widehat{\beta}_j)$ is the (jj) entry of the inverse of the observed Fisher information matrix.

Proportional hazards model

Semiparametric specification

↪ The (partial) log likelihood corresponding to the (partial) likelihood in (6) is given by

$$\log L_p(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left\{ \mathbf{x}_i' \boldsymbol{\beta} - \log \sum_{l \in R(y_i)} \exp(\mathbf{x}_l' \boldsymbol{\beta}) \right\}.$$

↪ The estimate of $\boldsymbol{\beta}$ is found by maximising this partial log likelihood function using numerical methods.

↪ This semiparametric proportional hazards model (sometimes simply termed as Cox model) is implemented in the `coxph` function in the `survival` package.

Proportional hazards model

Semiparametric specification

↪ In order to shed more light on the structure of the partial likelihood, let us consider the following toy dataset:

i	$y_i = \min\{t_i, c_i\}$	δ_i	x_i
1	9	1	4
2	8	0	5
3	6	1	7
4	10	1	3

↪ We now compile the pieces that go into the partial likelihood contributions at each event time. Note the ordered uncensored event times are 6, 9, and 10.

ordered event time	risk set	Likelihood contribution
6	$\{1,2,3,4\}$	$e^{7\beta}/(e^{7\beta} + e^{5\beta} + e^{4\beta} + e^{3\beta})$
9	$\{1,4\}$	$e^{4\beta}/(e^{4\beta} + e^{3\beta})$
10	$\{4\}$	$e^{3\beta}/e^{3\beta} = 1$

Proportional hazards model

Semiparametric specification

↪ The partial likelihood is the product of these terms, that is,

$$L(\beta) = \frac{e^{7\beta}}{e^{7\beta} + e^{5\beta} + e^{4\beta} + e^{3\beta}} \times \frac{e^{4\beta}}{e^{4\beta} + e^{3\beta}} \times 1$$
$$= \frac{e^{11\beta}}{(e^{7\beta} + e^{5\beta} + e^{4\beta} + e^{3\beta})(e^{4\beta} + e^{3\beta})}$$

↪ The estimate for β is the value, say $\widehat{\beta}$ that maximises this partial likelihood.

Proportional hazards model

Semiparametric specification

- ↪ Thus far, we have worked under the assumption that no ties are present among the (uncensored) event times, and thus, that the data can be uniquely sorted with respect to time.

- ↪ Although tied events are not possible in theory, given the continuous nature of T, they can however occur in practice.

- ↪ Event times are indeed often discrete in practice due to the fact that they are usually recorded to the nearest day, week, month, or year, depending on the application.

- ↪ The partial likelihood we have just learned about cannot be used in the case of tied event times. Kalbfleish and Prentice (2002) have proposed an appropriate expression of the partial likelihood for the case of ties, but this expression has a very complicated form and so it will not be reproduced here.

- ↪ For more details about how to handle ties, I refer to Collett (2014, chapter 3). For the matters of this course, it suffices to use the `coxph` function when analysing time to event data with ties under a Cox proportional hazards model (this function handles ties appropriately).

Proportional hazards model

Semiparametric specification

- Let us consider an example involving death times of male laryngeal cancer patients (example from Klein and Moeschberger, p 9).
- Kardaun (1983) reports data on 90 males diagnosed with cancer of the larynx during the period 1970–1978 at a Dutch hospital.
- Times recorded are the time (in years) between first treatment and either death or the end of the study (January, 1983).
- Also recorded are the patient's age at the time of diagnosis and the stage of the patient's cancer (Stage I to Stage IV). The stages are ordered from least serious to most serious.

Proportional hazards model

Semiparametric specification

- Let us use the following three indicator variables for modelling the effect of cancer stage on the hazard of death:
  - $X_1 = 1$ if the patient is in Stage II and 0 otherwise.
  - $X_2 = 1$ if the patient is in Stage III and 0 otherwise.
  - $X_3 = 1$ if the patient is in Stage IV and 0 otherwise.
- This places a patient with Stage I cancer in the baseline group, i.e., such a patient will have $X_1 = X_2 = X_3 = 0$.
- The proportional hazards model is then of the form

$$h(t \mid \mathbf{X}) = h_0(t) \exp(x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4),$$

  where X_4 encapsulates the age effect.
- Note that the baseline hazard $h_0(t)$ represents the hazard function for an individual for which $X_1 = X_2 = X_3 = X_4 = 0$, i.e., for an individual who is in Stage I and of zero years old! Here we are really not interested in the baseline hazard function, Cox semiparametric specification does not even rely on it, but in case we were, age would need to be centred!

Proportional hazards model

Semiparametric specification

- The estimates of the parameters are as follows:

$$\widehat{\beta}_1 = 0.14004, \quad \widehat{\beta}_2 = 0.64238, \quad \widehat{\beta}_3 = 1.70598, \quad \widehat{\beta}_4 = 0.01903.$$

- For instance, here $e^{10\widehat{\beta}_4} = 1.21$ represents the hazard ratio for an individual who is 10 years older than another one (e.g., 50 years against 40 years), with both being on the same disease stage. The hazard of death for an individual who is, say, 50 years old, is 1.21 times the hazard of death of an individual who is 40 years old. Both should be in the same disease stage group.

- Also, for two individuals of the same age, $e^{\widehat{\beta}_3} = 5.51$ gives the hazard ratio for an individual with stage IV cancer compared to the reference group (stage I). So, for two individuals of the same age, the hazard of death for an individual in stage IV of the disease is 5.5 that of a subject in stage I of the disease.

- Similar interpretations follow for $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

- Further, $e^{\widehat{\beta}_2 - \widehat{\beta}_1} = 1.65$ gives the hazard ratio for an individual with Stage III compared to an individual in Stage II.

Proportional hazards model

Semiparametric specification

- Similarly to what we have done when studying logistic regression models, we can also conduct a Wald test for testing the null hypothesis: $H_0 : \beta_j = 0$.

- As before,

$$\frac{\widehat{\beta}_j}{\widehat{\mathrm{SE}}(\widehat{\beta}_j)} \sim \mathrm{N}(0,1),$$

  or, equivalently,

$$\frac{\widehat{\beta}_j^2}{\widehat{\mathrm{var}}(\widehat{\beta}_j)} \sim \chi^2_1,$$

  under the null hypothesis H_0.

- Remember that this test however only considers one particular coefficient at a time, adjusted for the presence of the other risk factors in the model.

Proportional hazards model

Semiparametric specification

- Another alternative is to do a likelihood ratio test.
- If we want to compare two alternative models, say model A and B, where model A is nested in model B, the as before, we still have

$$2(\log L_B - \log L_A) \sim \chi^2_{p_B - p_A},$$

  where p_A and p_B denote the number of parameters in model A and model B, respectively.
- The AIC/BIC can also be used for model building.
- Please see the Supplementary Materials file for more details.

Proportional hazards model

Semiparametric specification: adjusting survival curves

- Since the baseline hazard is considered a nuisance parameter and it is not estimated, Cox semiparametric proportional hazards model cannot be used directly to obtain an estimator of the survival curves of subjects with specific covariate values.

- The baseline hazard has to be estimated and one proposal is to extend the Nelson–Aalen estimator for the case of risk factors:

$$\widehat{H}_0(t) = \sum_{j:t_j \leq t} \frac{d_j}{\sum_{l \in R(t_j)} \exp(\mathbf{x}_l' \widehat{\boldsymbol{\beta}})},$$

  with $t_1, \ldots, t_J$ being the ordered distinct event times and d_j the number of events at time t_j.

- The survival curve for subjects with covariate values $\mathbf{x}_i$ is then estimated by

$$\widehat{S}(t \mid \mathbf{x}_i) = \left[\widehat{S}_0(t)\right]^{\exp(\mathbf{x}_i' \widehat{\boldsymbol{\beta}})}, \quad \widehat{S}_0(t) = \exp\{-\widehat{H}_0(t)\}.$$

Proportional hazards model

Semiparametric specification: adjusting survival curves