Biostatistics (MATH11230)

Vanda Inácio

University of Edinburgh

Semester 1, 2021/2022

Proportional hazards model

↪ Last week we have learned about the proportional hazards model, which is given by

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

where $h_0(t)$ is the baseline hazard function and does not depend on the explanatory variables/risk factors.

↪ We have also seen that e^{β_j}, $j = 1, \ldots, p$, 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.

↪ 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.

Proportional hazards model

Fully parametric specification

- Choosing a parametric distribution for the event times 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

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

  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

- ↪ Estimates of the model parameters ($\boldsymbol{\beta}$ and the parameters of the parametric event times distribution, represented in the above likelihood by θ_1), can be obtained by maximising the likelihood function.

- ↪ The maximisation 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

- 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

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

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

↪ The probability that subject j has the event at time t given that one of the subjects from the risk set $R(t)$ had an event at time t is

$$\frac{\exp(\mathbf{x}_j'\boldsymbol{\beta})}{\sum_{l\in R(t)}\exp(\mathbf{x}_l'\boldsymbol{\beta})}.$$

↪ Because there are J event times and they are independent of one another we have

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

Proportional hazards model

Semiparametric specification

↪ 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 times times). Supposing that there are n times (censored and uncensored), and using the notation from slide 2, we have

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sum_{l \in R(y_i)} \exp(\mathbf{x}_l'\boldsymbol{\beta})} \right\}^{\delta_i} \tag{2}$$

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

↪ 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 expected Fisher information matrix.

- ↪ In practice, the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is approximated by the inverse of information matrix evaluated at $\boldsymbol{\beta}$.

- ↪ Based on the asymptotic normality, 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 approximated estimated variance-covariance matrix.

Proportional hazards model

Semiparametric specification

↪ The (partial) log likelihood corresponding to the (partial) likelihood in (2) 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 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	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.

ordered event time	risk set	Likelihood contribution
6	$\{1,2,3,4\}$	$e^{7\beta}/(e^{7\beta}+e^{5\beta}+e^{4\beta}+e^{3\beta})$
8	$\{1,2,4\}$	$\left[e^{5\beta}/(e^{5\beta}+e^{4\beta}+e^{3\beta})\right]^0 = 1$
9	$\{1,4\}$	$e^{4\beta}/(e^{4\beta}+e^{3\beta})$
10	$\{4\}$	$e^{3\beta}/e^{3\beta} = 1$

↪ The partial likelihood is the product of these four terms.

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

Proportional hazards model

Semiparametric specification

- ↪ Thus far, we have worked under the assumption that no ties are present among the 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 intervals (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 referent 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.

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).

- 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.

- Please see the Supplementary Materials file for more details.