

Lecture 28 — The proportional hazards model

28.1 The proportional hazards model

Example 28.1. A clinical trial is performed to study the effect of a drug for maintaining/prolonging remission induced by chemotherapy in the treatment of acute leukemia. (Remission is the disappearance of leukemic cells and other symptoms of the disease.) For each i^{th} patient in the trial, let T_i denote the length of the remission (or equivalently, the time until recurrence of the cancer), which we wish to model in terms of patient-specific covariates x_{i1}, \dots, x_{ip} . The first covariate x_{i1} may be a 0–1 variable indicating whether patient i received the drug or a placebo, and the remaining covariates are other factors, such as the age of the patient, that may affect the remission length.

Modeling T_i as a continuous, positive-valued random variable with CDF $F_i(t)$ and PDF $f_i(t) = F_i'(t)$, it is useful to think about the distribution of T_i in terms of its **hazard function** $\lambda_i(t)$, which represents the “instantaneous risk” of recurrence at time t :

$$\lambda_i(t) := \lim_{\delta \rightarrow 0} \frac{1}{\delta} \mathbb{P}[T_i \leq t + \delta \mid T_i \geq t].$$

In other words, for small δ , the probability that a recurrence of the cancer occurs in the time window $[t, t + \delta]$, conditional on it not having occurred up to time t , is approximately $\delta \lambda_i(t)$. The hazard function may be expressed in terms of the CDF $F_i(t)$ and PDF $f_i(t)$ as

$$\lambda_i(t) = \lim_{\delta \rightarrow 0} \frac{\mathbb{P}[t \leq T_i \leq t + \delta]}{\delta \mathbb{P}[T_i \geq t]} = \lim_{\delta \rightarrow 0} \frac{F_i(t + \delta) - F_i(t)}{\delta(1 - F_i(t))} = \frac{f_i(t)}{1 - F_i(t)}.$$

To develop some intuition for the hazard function, consider a simple example where $T_i \sim \text{Exponential}(\theta)$. Then the PDF is $f_i(t) = \theta e^{-\theta t}$, the CDF is $F_i(t) = 1 - e^{-\theta t}$, so the hazard function is

$$\lambda_i(t) = \frac{\theta e^{-\theta t}}{1 - (1 - e^{-\theta t})} = \theta.$$

In this case, the hazard function is constant in time (which is a special property of the exponential distribution). Intuitively, this means that assuming the remission has lasted until time t , the probability of the recurrence occurring in the next instant of time is the same for every t and is determined only by θ . The parameter θ governs how quickly the exponential distribution decays—the larger the value of θ , the faster the rate of decay, and the more likely it is that recurrence of the cancer will occur at any next instant of time.

Cox’s **proportional hazards model** does not assume that the distribution of T_i is exponential, or that it follows any particular parametric form. Instead, it models the hazard function for T_i as

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 x_{i1} + \dots + \beta_p x_{ip}).$$

The regression coefficients β_1, \dots, β_p are unknown parameters determining the effects of the covariates on the remission length T_i , and $\lambda_0(t)$ is a completely unknown **baseline hazard function**. In this model, $\lambda_0(t)$ controls the shape of the hazard function over time for *all* patients, and the factor $\exp(\beta_1 x_{i1} + \dots + \beta_p x_{ip})$ controls the scale of the hazard function for each patient i . Thus the model asserts that for any two patients i and j , their hazard functions have the same shape and differ only in scale, so that the ratio of their hazard functions $\lambda_i(t)/\lambda_j(t)$ is constant over time (hence the name “proportional” hazards). The model posits that this scale ratio is determined by a linear combination of the differences of the patients’ covariates.

In the clinical trial, the remission for a patient i may last longer than the duration for which the patient participates in the trial. In this case, we do not observe their true remission length T_i (which can take the value ∞ if the cancer never returns), but instead we only observe that $T_i > l_i$ where l_i is the length of time for which the patient is in the trial. This type of observation is called **right-censored**. The method of inference developed below for the proportional hazards model will naturally handle data in which some of the observations are right-censored. We treat l_i as a fixed and known constant for every patient, so that we either observe a value of T_i that is at most l_i , or we observe that $T_i > l_i$. We will make an important assumption that T_i (the true remission length) does not depend on l_i (the right-censoring time).

The proportional hazards model may be used to model the time-to-onset of any event pertaining to an individual in terms of observed covariates for that individual; example applications include medical trials as above, as well as industrial reliability experiments that model the time-to-failures of devices. According to a 2014 list in the scientific journal *Nature*, the 1972 paper by David Cox which introduced this model is the 2nd most cited paper in statistics and the 24th most cited paper in all of science.

28.2 Statistical inference

In many applications, the regression coefficients β_1, \dots, β_p are of greater interest than the baseline hazard function $\lambda_0(t)$. If the first covariate x_{i1} corresponds to an indicator variable representing assignment to the treatment group (drug) or the control group (placebo), then the coefficient β_1 represents the log-hazards-ratio between the two groups after controlling for the other covariates x_{i2}, \dots, x_{ip} . We will discuss inference procedures for the following tasks:

- Estimate β_1, \dots, β_p .
- Test whether $\beta_1 = 0$.

Perhaps surprisingly, it is possible to perform these inference tasks without any knowledge of, and without any assumptions regarding, the baseline hazard function $\lambda_0(t)$.

In previous models, we performed inference by writing down the likelihood of the model parameters. Inference in the proportional hazards model will be slightly different from these previous examples, because the baseline hazard function $\lambda_0(t)$ is completely unknown, and the likelihood function and MLEs for β_1, \dots, β_p would depend on $\lambda_0(t)$. If $\lambda_0(t)$ were modeled

parametrically using a small number of additional parameters, then we may include these as parameters of the model and fit the entire model by computing the joint MLEs of these additional parameters and β_1, \dots, β_p . However, without parametric modeling assumptions on $\lambda_0(t)$, in this course we have not discussed procedures for how to estimate an entire unknown function $\lambda_0(t)$.

We will circumvent this problem by conditioning on the set of all distinct observed recurrence times $t_{(1)} < t_{(2)} < \dots < t_{(m)}$ across all patients. (This idea is quite similar to how we conditioned on the set of all distinct observed values in permutation two-sample tests, to address the problem of an unknown common distribution function $F = G$ under the null hypothesis.) Since we are modeling T_i as continuous random variables, we may assume that each observed recurrence time $t_{(k)}$ corresponds to only one patient (i.e. there are no ties in recurrence times), so m is just the total number of non-right-censored observations. For each $t_{(k)}$, the *risk set* $\mathcal{R}_{(k)}$ immediately before time $t_{(k)}$ is the set of patients who have not yet left the study (been right-censored) and are still in remission—this represents the candidate set of patients for which we may have observed a recurrence at time $t_{(k)}$. Conditional on the fact that some patient in this risk set $\mathcal{R}_{(k)}$ has a recurrence at time $t_{(k)}$, the probability that it is a particular patient $I_k \in \mathcal{R}_{(k)}$ is

$$\frac{\lambda_{I_k}(t_{(k)})}{\sum_{i \in \mathcal{R}_{(k)}} \lambda_i(t_{(k)})}$$

(the ratio of the “instantaneous rate” of recurrence for patient I_k to the sum of the rates for all candidate patients). Under the proportional hazards model, this is

$$\frac{\lambda_0(t_{(k)}) \exp(\beta_1 x_{I_k 1} + \dots + \beta_p x_{I_k p})}{\sum_{i \in \mathcal{R}_{(k)}} \lambda_0(t_{(k)}) \exp(\beta_1 x_{i1} + \dots + \beta_p x_{ip})}.$$

Importantly, the factor $\lambda_0(t_{(k)})$ cancels from the numerator and denominator of this expression, yielding a quantity that does not depend on the baseline hazard function $\lambda_0(t)$. Taking a product of the above expression over all observed recurrence times yields

$$\text{plik}(\beta_1, \dots, \beta_p) = \prod_{k=1}^m \frac{\exp(\beta_1 x_{I_k 1} + \dots + \beta_p x_{I_k p})}{\sum_{i \in \mathcal{R}_{(k)}} \exp(\beta_1 x_{i1} + \dots + \beta_p x_{ip})}.$$

This quantity is called the **partial likelihood function** of β_1, \dots, β_p . Intuitively, it captures all of the information contained by the observations that at each time $t_{(k)}$, the particular recurrence was for patient I_k as opposed to the other patients for which we could have observed a recurrence at that time. We may perform likelihood-based inference using this partial likelihood in place of the usual likelihood function.

We may estimate β_1, \dots, β_p by maximizing the partial likelihood over these parameters. As with usual MLE calculations, it is computationally convenient to first take a logarithm, so we consider the log-partial likelihood

$$l(\beta_1, \dots, \beta_p) = \sum_{k=1}^m \left(\beta_1 x_{I_k 1} + \dots + \beta_p x_{I_k p} - \log \sum_{i \in \mathcal{R}_{(k)}} \exp(\beta_1 x_{i1} + \dots + \beta_p x_{ip}) \right).$$

We may maximize this quantity by setting its derivative with respect to each β_1, \dots, β_p equal to 0:

$$0 = \frac{\partial l}{\partial \beta_j} = \sum_{k=1}^m \left(x_{I_{kj}} - \frac{\sum_{i \in \mathcal{R}_{(k)}} x_{ij} \exp(\beta_1 x_{i1} + \dots + \beta_p x_{ip})}{\sum_{i \in \mathcal{R}_{(k)}} \exp(\beta_1 x_{i1} + \dots + \beta_p x_{ip})} \right).$$

Solving numerically this system of p equations in p unknowns β_1, \dots, β_p yields the maximum partial-likelihood estimates $\hat{\beta}_1, \dots, \hat{\beta}_p$.

The asymptotic theory for the maximum partial-likelihood estimate is analogous to that of the MLE in usual parametric models (although the mathematical derivation of this theory requires more advanced probabilistic tools that we did not cover in this course). In particular, the usual generalized likelihood ratio test applies: To test $H_0 : \beta_1 = 0$, we may compute the maximum partial likelihood estimates $\hat{\beta}_{2,0}, \dots, \hat{\beta}_{p,0}$ in this sub-model, using the same procedure as above with the first covariate removed. The test statistic

$$-2 \log \Lambda = -2 \log \frac{\text{plik}(0, \hat{\beta}_{2,0}, \dots, \hat{\beta}_{p,0})}{\text{plik}(\hat{\beta}_1, \dots, \hat{\beta}_p)}$$

is, under mild regularity conditions, distributed as χ_1^2 in the limit of large n , and an asymptotic level- α test would reject H_0 when $-2 \log \Lambda$ exceeds the upper- α point $\chi_1^2(\alpha)$.