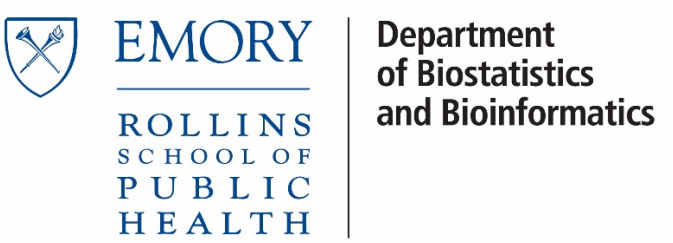


BIOS 522: Survival Analysis Methods

Reading 5:

Introduction to the Cox model

This week, we will review linear and logistic regression. We will introduce Cox proportional hazards regression for modeling time-to-event data. We will define the partial likelihood and how to handle tied survival times. We will predict survival under the Cox proportional hazards model.

Part 1. Review of linear and logistic regression

So far in this course we have considered the **one-sample problem** of studying a single population (fitting a Kaplan-Meier curve). We have also considered the **two-sample problem** of comparing two independent populations (visually comparing Kaplan-Meier curves; running a log-rank test). For more complex settings (multiple covariates, continuous covariates, interactions), we turn to **regression**.

Linear regression

The simplest regression model is **linear regression**. Let Y_i denote a continuous response variable for individual i, and let $X_{i1}, \ldots, X_{ik}$ be a set of k covariates that we wish to use to predict Y_i. These covariates can be either categorical or continuous. Then a linear regression model has the form:

$$E[Y_i] = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

The **intercept** β_0 can be interpreted as the expected/mean value of Y_i for the population with all covariates equal to zero. Sometimes this population is referred to as the **reference population**, but this population may not always exist or be meaningful (e.g. blood pressure for people with BMI equal to zero).

Each coefficient β_j can be interpreted as the change in the expected/mean value of Y_i for a one-unit increase in X_{ij} when all other covariates are held constant. It is commonly referred to as the **slope**. For a binary covariate, β_j is the difference in means of Y_i for the group with $X_{ij} = 1$ compared to the group with $X_{ij} = 0$, when all other covariates are held constant.

We can fit the model using least squares and obtain estimates of the parameters $\hat{\beta}_0, \ldots, \hat{\beta}_k$. If we wish to obtain a predicted value $\hat{Y}_i$ for person i with covariates $X_{i1}, \ldots, X_{ik}$, we can enter their values into the equation:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_k X_{ik}$$

Often, we use regression models to make inference about the relationships between different covariates and the response variable. We may be interested in testing the null hypothesis that a covariate X_{ij} has no effect on the response variable Y_i after adjusting for all other covariates; this is equivalent to testing H_0: $\beta_j = 0$. We can construct a **Wald test** of H_0 from $\hat{\beta}_j$ and standard error $\widehat{SE}(\hat{\beta}_j)$:

$$Z = \frac{\hat{\beta}_j - 0}{\widehat{SE}(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)}$$

Under H_0, the test statistic Z approximately follows a standard normal distribution. It is significant at the α =0.05 level if $Z > 1.96$ or $Z < -1.96$.

Logistic regression

Another frequently used model is **logistic regression**. Let Y_i denote a binary (yes/no) response variable for individual i, and let $X_{i1}, \ldots, X_{ik}$ be a set of k covariates. Let p_i be the probability that individual i will have a positive response ($Y_i = 1$). Then a logistic regression model has the form:

$$\text{logit}[p_i] = \log[p_i/(1 - p_i)] = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

The **intercept** β_0 can be interpreted as the **log odds** of Y_i in the population with all covariates equal to zero. As before, this reference population may not be meaningful.

Each coefficient β_j can be interpreted as the **log odds ratio** of Y_i associated with a one-unit increase in X_{ij} when all other covariates are held constant. For a binary covariate, β_j is the log odds ratio of Y_i for the group with $X_{ij} = 1$ compared to the group with $X_{ij} = 0$, adjusting for all other covariates.

We fit the model using maximum likelihood and obtain estimates of these parameters $\hat{\beta}_0, \ldots, \hat{\beta}_k$. If we wish to obtain a predicted probability of positive response $\hat{p}_i$ for person i with covariates $X_{i1}, \ldots, X_{ik}$, we can enter their values into the following equation:

$$\hat{p}_i = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_k X_{ik}\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_k X_{ik}\right)}$$

We may be interested in testing the null hypothesis that a covariate X_{ij} has no effect on the probability of positive response p_i after adjusting for all other covariates. If there is no effect, the odds ratio is equal to 1, or equivalently, the log odds ratio β_j is equal to 0. We can construct a Wald test for H_0: $\beta_j = 0$ from $\hat{\beta}_j$ and its standard error $\widehat{SE}(\hat{\beta}_j)$:

$$Z = \frac{\hat{\beta}_j - 0}{\widehat{SE}(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)}$$

Under H_0, the test statistic Z approximately follows a standard normal distribution. It is significant at the α =0.05 level if $Z > 1.96$ or $Z < -1.96$.

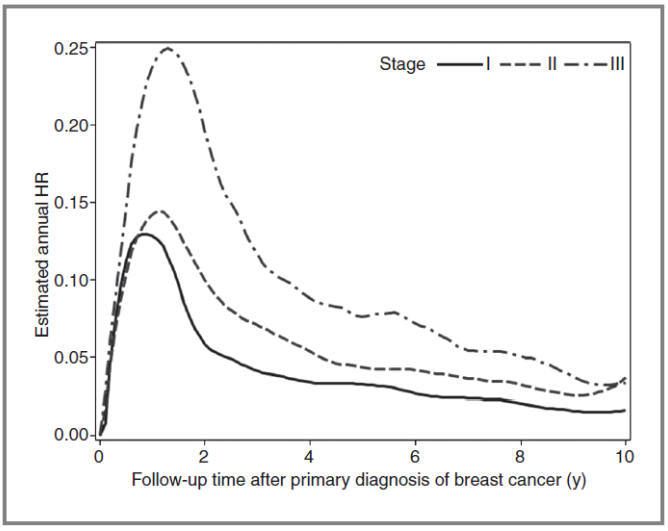
Part 2. Basics of the Cox proportional hazards model

Motivation

The log-rank test can be used to compare survival times for two or more different groups. However, it cannot be used if we are interested in the relationship between survival time and a continuous covariate, or to evaluate the simultaneous effects of more than one covariate. And while it returns a p-value for the statistical significance of the comparison, it does not provide a point estimate summarizing the *magnitude* of the covariate's effect on survival.

The **Cox proportional hazards regression model** is the most commonly used regression model for time-to-event data. It accommodates right-censoring. The general structure of Cox proportional hazards regression model resembles that of other regression models. The primary difference is that our dependent variable is now the *hazard function* $h(t)$ for all times t. We model the effect of covariates through their impact on the entire hazard function.

Consider the hazard functions for first recurrence among women after primary breast cancer treatment for stage I, II, and III cancer (Cheng et al. 2012). We see in the following figure that the period of highest hazard is between 1 and 3 years after primary breast cancer treatment. Women with stage III cancer have an elevated hazard at all times compared to women with stage II cancer.



We can examine the ratio of these hazard functions. At all times, the hazard for women with stage III breast cancer is approximately double the hazard for women with stage II breast cancer. We refer to this ratio as the **hazard ratio**. In this example, the hazard functions are roughly **proportional** because the hazard ratio is always around 2 at all times t.

We can imagine modeling the hazard function for all groups as (i) the baseline hazard of a reference group (e.g. patients with stage II cancer), and (ii) a multiplicative term (the hazard ratio) which defines how much higher or lower each group's hazard function is than that of the reference group.

Proportional hazards model formulation

Let $h_i(\cdot)$ be the hazard function for individual i. Let $X_{i1}, \dots, X_{ik}$ be a set of k covariates that we wish to use to model $h_i(\cdot)$. These covariates can be either categorical or continuous. Then a Cox proportional hazards regression model has the form:

$$h_i(\cdot) = h_0(\cdot)\exp(\beta_1 X_{i1} + \cdots + \beta_k X_{ik})$$

An individual's hazard is modeled as the product of two terms. The first term, $h_0(\cdot)$, is referred to as the **baseline hazard**. The second term, $\exp(\beta_1 X_{i1} + \cdots + \beta_k X_{ik})$, is a hazard ratio that captures multiplicatively how much higher or lower that individual's hazard will be when compared to the baseline level, depending on the values of their covariates.

The baseline hazard can be interpreted as the hazard function for the reference group, here the group with all covariates equal to 0.

$$\begin{aligned} h_i(\cdot) &= h_0(\cdot)\exp(\beta_1 \times 0 + \cdots + \beta_k \times 0) \\ &= h_0(\cdot)\exp(0) \\ &= h_0(\cdot) \end{aligned}$$

The baseline hazard takes the place of an intercept term. As before, this reference population may not be meaningful.

A key advantage of the Cox proportional hazards model is that we do not need to specify a parametric distribution for $h_0(\cdot)$. Thus, the baseline hazard can take any shape and can vary over time. It can increase and then decrease like in the breast cancer data example above.

Each coefficient β_j can be interpreted as the **log hazard ratio** of T_i associated with a one-unit increase in X_{ij} when all other covariates are held constant. Thus, $\exp(\beta_j)$ is the **hazard ratio**. Consider person i with $X_{i1}, \dots, X_{ij}, \dots, X_{ik}$, and consider person i' with covariates $X_{i1}, \dots, (X_{ij}+1), \dots, X_{ik}$. Then, at any time t:

$$\begin{aligned}
\frac{h_i'(t)}{h_i(t)} &= \frac{h_0(t)\exp\left(\beta_1 X_{i1} + \cdots + \beta_j\left(X_{ij}+1\right) + \cdots + \beta_k X_{ik}\right)}{h_0(t)\exp\left(\beta_1 X_{i1} + \cdots + \beta_j X_{ij} + \cdots + \beta_k X_{ik}\right)} \\
&= \exp\left[\left(\beta_1 X_{i1} + \cdots + \beta_j\left(X_{ij}+1\right) + \cdots + \beta_k X_{ik}\right)\right. \\
&\qquad \left. - \left(\beta_1 X_{i1} + \cdots + \beta_j X_{ij} + \cdots + \beta_k X_{ik}\right)\right] \\
&= \exp\left[\beta_j\left(X_{ij}+1\right) - \beta_j X_{ij}\right] \\
&= \exp\left(\beta_j\right) \\
\log\left[\frac{h_i'(t)}{h_i(t)}\right] &= \beta_j
\end{aligned}$$

Note how the baseline hazard function cancels in the above. Furthermore, note how this ratio is equal to $\exp(\beta_j)$ at all times t.

For a binary covariate, β_j is the log hazard ratio of Y_i for the group with $X_{ij} = 1$ compared to the group with $X_{ij} = 0$, holding all other covariates constant.

- If $\beta_j > 0$, the hazard ratio $\exp(\beta_j) > 1$; this means that a higher value of X_{ij} is associated with a higher hazard and worse survival.
- If $\beta_j < 0$, the hazard ratio $\exp(\beta_j) < 1$; this means that a higher value of X_{ij} is associated with a lower hazard and better survival.
- If $\beta_j = 0$, the hazard ratio $\exp(\beta_j) = 1$; this means that different values of X_{ij} will have the same hazard and same survival.

The hazard ratio on the log scale

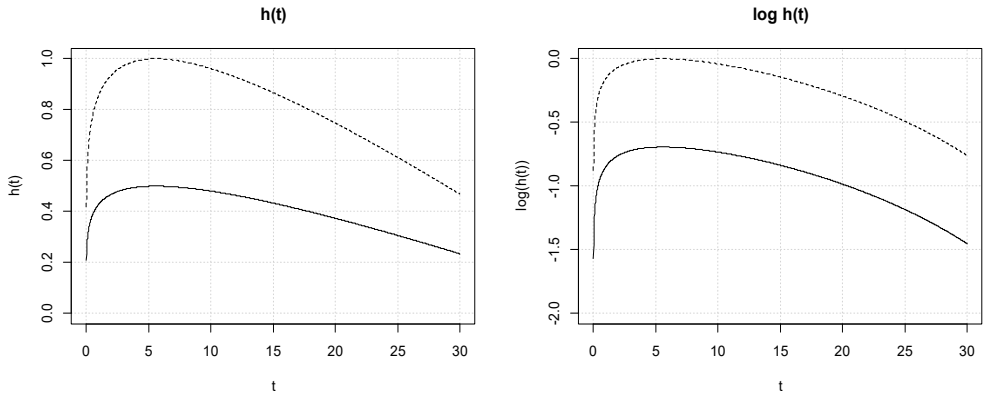
It can be particularly helpful to view proportional hazards on the log scale. Cox regression models assume that the effects of covariates on the hazard function are multiplicative. Equivalently, the effects of covariates on the log hazard function are additive.

$$\begin{aligned}
\log\left(h_i(\cdot)\right) &= \log\{h_0(\cdot)\exp(\beta X_i)\} \\
&= \log\left(h_0(\cdot)\right) + \log(\exp(\beta X_i))
\end{aligned}$$

$$= \log\big(h_0(\cdot)\big) + \beta X_i$$

Thus, in the Cox model, the difference between the log hazard function for the group with $X_i = 1$ and the log hazard function for the group with $X_i = 0$ is the addition of a constant value β.

A figure is provided below to visualize the hazard functions on both scales. In the left figure, we see two hazard functions. Their values vary over time, but their ratio is always constant (the dotted line is always twice as high as the solid line). In the right figure, we see these same hazard functions plotted on the log scale. Again, their values vary over time, but the lines are parallel. The absolute difference between these is always constant (the dotted line is always 0.69, i.e. log(2), higher than the dashed line).



We can examine the cumulative hazard function under the proportional hazards model:

$$\begin{aligned} H_i(t) &= \int_{u=0}^{u=t} h_i(u)du \\ &= \int_{u=0}^{u=t} h_0(u)\exp(\beta X_i)\,du \\ &= \exp(\beta X_i)\int_{u=0}^{u=t} h_0(u)du \\ &= \exp(\beta X_i)\,H_0(t) \end{aligned}$$

We can see that the log cumulative hazard functions behave similarly to the log hazard functions:

$$\begin{aligned} \log\big(H_i(t)\big) &= \log\{H_0(t)\exp(\beta X_i)\} \\ &= \log\big(H_0(t)\big) + \log(\exp(\beta X_i)) \\ &= \log\big(H_0(t)\big) + \beta X_i \end{aligned}$$

Thus, under the proportional hazards model, the log cumulative hazard functions are also parallel.

Part 3. Semiparametric regression and the partial likelihood

Semiparametric regression

The Cox proportional hazards model has the following form:

$$h_i(\cdot) = h_0(\cdot)\exp(\beta_1 X_{i1} + \cdots + \beta_k X_{ik})$$

Cox's model is known as a **semiparametric regression model** because it has parametric component and a nonparametric component. The nonparametric component of the model is the flexible baseline hazard $h_0(\cdot)$. The parametric component of the model is the term $\exp(\beta_1 X_{i1} + \cdots + \beta_k X_{ik})$, where each covariate X_{ij} has a multiplicative effect on the hazard function.

In general, we are most interested in estimating the log hazard ratios $\beta_1, \dots, \beta_k$ as these allow us to summarize the effect of our covariates on survival. We are less interested in estimating the infinitely dimensional baseline hazard $h_0(\cdot)$, in part because it is more complex to fit.

We are able to estimate the $\beta_1, \dots, \beta_k$ without estimating the entire baseline hazard $h_0(\cdot)$. This is because the full likelihood can be split into two terms:

$$L(\beta_1, \dots, \beta_k)L(h_0(\cdot), \beta_1, \dots, \beta_k)$$

The first term $L(\beta_1, \dots, \beta_k)$ depends on $\beta_1, \dots, \beta_k$ but does not depend on $h_0(\cdot)$. It is known as the **partial likelihood** because it is only one part of the full likelihood. The second term depends on $h_0(\cdot)$, as well as containing a little bit of information on the log hazard ratios.

In Cox proportional hazards regression, instead of calculating the full likelihood for the data, we calculate the partial likelihood. We find the values of $\beta_1, \dots, \beta_k$ where the partial likelihood maximizes (as if it were a full likelihood). These are our **maximum partial likelihood estimates** $\hat{\beta}_1, \dots, \hat{\beta}_k$. The maximum partial likelihood estimator is consistent and asymptomatically normal.

The Cox partial likelihood formulation

The form of the Cox partial likelihood is a product over unique failure times in the data set. Suppose there is a failure in person i at time T_i, and let $R(T_i)$ denote the set of individuals at risk at time T_i (the **risk set**). The partial likelihood contribution is the conditional probability that, given that there was a failure at time T_i with risk set $R(T_i)$, that individual i (with their unique combination of covariates) was the one to fail.

For simplicity, assume that there is one unknown β and one covariate X_j for each person j. The partial likelihood contribution is:

$$\frac{h_i(T_i)}{\sum_{j\in R(T_i)} h_j(T_i)} = \frac{h_0(T_i)\exp(\beta X_i)}{\sum_{j\in R(T_i)} h_0(T_i)\exp(\beta X_j)} = \frac{\exp(\beta X_i)}{\sum_{j\in R(T_i)}\exp(\beta X_j)}$$

The partial likelihood contribution will be highest when the Cox model predicts that individual i's covariates put them at elevated risk compared to the others still at risk. Note that the baseline hazard cancels out, so the model does not need to make any assumptions about it.

The partial likelihood is the product of these contributions over all failure times $T_1, \ldots, T_m$:

$$L(\beta) = \prod_{i=1}^{m} \frac{\exp(\beta X_i)}{\sum_{j \in R(T_i)} \exp(\beta X_j)}$$

For estimation of β, it is maximized like a full likelihood. If there are multiple covariates in the model, $\exp(\beta X_i)$ is replaced by

$$\exp(\beta_1 X_{i1} + \cdots + \beta_k X_{ik})$$

where X_{ij} is the j^{th} covariate for person i.

Handling ties

The Cox partial likelihood is calculated by examining each unique failure time and comparing the covariates of the individual who failed with those at risk. The conditional probability calculation assumes that all failure times are unique, i.e., there are no tied failure times. This is sensible for an underlying continuous random variable. In practice, tied failure times occur because of imperfect observation and rounding.

There are three common methods for dealing with ties in a Cox model (in order of complexity): Breslow, Efron, and exact. The Breslow approximation is the simplest but least accurate. The exact method is accurate but computationally complex. The Efron approximation is accurate and computationally efficient.

If there are d_i tied survival times at T_i, the **exact method** calculates the mean partial likelihood contribution at time T_i over all $d_i!$ possible ways of breaking ties among the people who failed at time T_i. These calculations get complex quickly – there are $5! = 120$ ways to break ties among 5 events, $10! =$ 3,628,800 ways among 10 events, and $15! = 1.3$ trillion ways among 15 events. For simplicity, we illustrate the calculation for two events.

Let y and z denote the indices of individuals who fail at time T_i, with $d_i = 2$. Let X_y and X_z denote their covariates. There are two ways to break the tie: y fails first, or z fails first. If y fails first, then the partial likelihood contribution from the two failures is:

$$\frac{\exp(\beta X_y)}{\sum_{j \in R(T_i)} \exp(\beta X_j)} \frac{\exp(\beta X_z)}{\sum_{j \in R(T_i) \setminus \{y\}} \exp(\beta X_j)}$$

Where $R(T_i)\backslash\{y\}$ denotes the risk set at T_i with y removed. We can calculate a similar partial likelihood contribution for the setting where z fails first. The mean partial likelihood contribution over these two possibilities is:

$$\frac{1}{2}\left[\frac{\exp(\beta X_y)}{\sum_{j\in R(T_i)}\exp(\beta X_j)}\frac{\exp(\beta X_z)}{\sum_{j\in R(T_i)\backslash\{y\}}\exp(\beta X_j)}\right]+\frac{1}{2}\left[\frac{\exp(\beta X_z)}{\sum_{j\in R(T_i)}\exp(\beta X_j)}\frac{\exp(\beta X_y)}{\sum_{j\in R(T_i)\backslash\{z\}}\exp(\beta X_j)}\right]$$

Extending this to larger numbers is straightforward but tedious. When there are failure times with large numbers of ties, the quickly becomes computationally intractable.

Note that in the above exact calculation, the numerators of each term are the same, but the denominators are slightly different. The **Efron approximation** approximates the mean denominator of the terms in the exact estimate. It does this by subtracting the average amount of time the observation is removed from the risk set for each of the tied failures.

Let D_i denote the set of people who fail at time T_i, so the set D_i contains d_i indices. The Efron approximation to the exact mean partial likelihood contribution is:

$$\frac{\prod_{j\in D_i}\exp(\beta X_j)}{\prod_{k=1}^{d_i}\left[\sum_{j\in R(T_i)}\exp(\beta X_j)-\frac{k-1}{d_i}\sum_{j\in D_i}\exp(\beta X_j)\right]}$$

This can be readily calculated and is a very good approximation to the exact method even when there are many ties.

In our example with two tied failure times, the Efron approximation to the likelihood contribution is:

$$\frac{\exp(\beta X_y)\exp(\beta X_z)}{\sum_{j\in R(T_i)}\exp(\beta X_j)\left[\sum_{j\in R(T_i)}\exp(\beta X_j)-\frac{1}{2}\left(\exp(\beta X_y)+\exp(\beta X_z)\right)\right]}$$

The **Breslow approximation** is the simplest, ignoring the fact that failures are removed from the risk set when breaking ties. This yields the following approximation to the exact partial likelihood contribution:

$$\frac{\prod_{j\in D_i}\exp(\beta X_j)}{\left[\sum_{j\in R(T_i)}\exp(\beta X_j)\right]^{d_i}}$$

In our example with tied failure times in y and z, the Breslow approximation to the exact partial likelihood contribution is:

$$\frac{\exp(\beta X_y)\exp(\beta X_z)}{\left[\sum_{j\in R(T_i)}\exp(\beta X_j)\right]^2}$$

These tie-breaking methods are available as options in statistical software.

Part 3. Predicted survival

Predicting survival probabilities

Fitting the Cox proportional hazards model gives us estimates of the log hazard ratios associated with the covariates in the model. Imagine that we also had an estimate $\widehat{H}_0(t)$ of the baseline cumulative hazard function. if we wish to predict a person's survival probability at time t, we can enter their covariates $X_{i1}, \ldots X_{ik}$ into the following equation:

$$\hat{S}_i(t) = \left[\exp\left(-\widehat{H}_0(t)\right)\right]^{\exp(\widehat{\beta}_1 X_{i1}+\cdots+\widehat{\beta}_k X_{ik})}$$

We will prove to ourselves that this works in this week's activity.

Recall that $\exp\left(\hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_k X_{ik}\right)$ is the estimated hazard ratio comparing an individual with covariates $X_{i1}, \ldots, X_{ik}$ to the reference group. Notice that the covariates control whether an individual is expected to fail *earlier* or *later* than the reference group. This is achieved through the hazard ratio term.

How might we estimate the baseline cumulative hazard function?

The above does not explain *how* we estimate $\widehat{H}_0(t)$.

One approach might be to fit a Nelson-Aalen estimator to just individuals in the data in the reference group, but there are major disadvantages to that:

- The reference group may not exist in the data, i.e., no observation with all covariates equal to 0.
- Where it does exist, the reference group could be quite small.

A better approach is to find an estimator that uses the *entire data set* to get a smoother estimate of the baseline hazard function or the baseline cumulative hazard function. There are several methods described in the literature.

Methods for estimating the baseline cumulative hazard function

In the absence of ties, estimation of $H_0(t)$ is based on the true likelihood function. The estimated hazard function $\hat{h}_0(t) = 0$ except for times at which a failure occurs. The profile maximum likelihood of $h_{0j} = h_0\left(T_j\right)$ is given by:

$$\hat{h}_{0j} = \frac{1}{\sum_{i \in R(T_j)} \exp\left(\hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_k X_{ik}\right)}$$

where $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are the maximum partial likelihood estimates of $\beta_1, \ldots, \beta_k$. An estimate of the baseline cumulative hazard function $H_0(t)$ is given by:

$$\widehat{H}_0(t) = \sum_{j:t_j \leq t} \frac{1}{\sum_{i \in R(T_j)} \exp\left(\hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_k X_{ik}\right)}$$

In the presence of ties, a related approach to estimating the baseline cumulative hazard function is the **Breslow estimator**, given by:

$$\widetilde{H}_0(t) = \sum_{j:t_j \leq t} \frac{d_j}{\sum_{i \in R(T_j)} \exp\left(\hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_k X_{ik}\right)}$$

The Breslow estimator of the baseline cumulative hazard function is closely related to the Nelson-Aalen estimator of the cumulative hazard function. If we imagine fitting a Cox model where all coefficients are zero, the Breslow estimator reduces to the Nelson-Aalen estimator fit to the entire data set.

$$\begin{aligned} \widetilde{H}_0(t) &= \sum_{j:t_j \leq t} \frac{d_j}{\sum_{i \in R(T_j)} \exp(0)} \\ &= \sum_{j:t_j \leq t} \frac{d_j}{n_j} \end{aligned}$$

Other methods include the **Kalbfleisch/Prentice estimator**, which is related to the Kaplan-Meier estimator, and the **Efron estimator**.

Regardless of how the baseline cumulative hazard function is calculated, we can use the following equation to predict survival probabilities for any combination of covariates:

$$\left[e^{-\widetilde{H}_0(t)}\right]^{\exp\left(\widehat{\beta}_1 X_{i1} + \cdots + \widehat{\beta}_k X_{ik}\right)}$$

Part 4. Looking ahead

In the coming weeks, we will continue to discuss the Cox model.

Lecture 6

- Interpreting hazard ratios for binary and continuous data
- Interval estimation and hypothesis testing

- Interactions, transformations, polynomials, splines
- Summarizing Cox model results

Lecture 7

- Examining the proportional hazards assumption
- Residuals
- Goodness of fit
- Visual diagnostics
- Model selection

Lecture 8

- Time-dependent covariates
- Time-varying effects
- Stratified Cox model