Survival Models

Max Welz
welz@ese.eur.nl

Econometric Institute
Erasmus School of Economics

October 11, 2021

1 Essential Theory

1.1 Setup

Suppose we want to model the failure rate of a some statistical process. Concretely, let the random variable T denote the (non-negative) real-valued failure time of the process of interest. We assume that the process fails eventually and that we observe it infinitely long so that we can observe its eventual failure time T (this assumption will be relaxed later). Let F be the distribution function of T and f be its corresponding density. By definition, for some time $t \in [0, \infty]$, the distribution function

$$F(t) = \mathbb{P}[T \leq t] = \int_0^t f(s)\mathrm{d}s$$

measures the probability that the process fails before or at time t. We call F the *incidence function*. The survival function S is defined by

$$S(t) = 1 - F(t) = \mathbb{P}[T > t],$$

and denotes the probability that failure occurs *after* time t. An essential quantity in survival modeling is the *hazard function* $h : [0, \infty] \to [0, \infty)$, defined by

$$h(t) = \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}[t \leq T < t + \Delta t | T \geq t]}{\Delta t}. \tag{1}$$

The hazard function $h(t)$ is interpreted as the instantaneous rate of failure, provided that the process has not yet failed. We emphasize that the hazard function is *not* a probability, which is a frequent misconception.

Proposition 1.1. *Let random variable T denote a failure time, let F and f its incidence function and corresponding density, respectively, and h its hazard function. It holds that*

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}.$$

Proof. Since f is the derivative of F, we can write

$$\begin{aligned} f(t) &= \lim_{\Delta t \downarrow 0} \frac{F(t+\Delta t) - F(t)}{\Delta t} = \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}[T \leq t + \Delta t] - \mathbb{P}[T \leq t]}{\Delta t} \\ &= \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}[t \leq T < t + \Delta t]}{\Delta t}. \end{aligned}$$

Thereupon, we can write the hazard function h as

$$\begin{aligned} h(t) &= \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}[t \leq T < t + \Delta t | T \geq t]}{\Delta t}. \\ &= \frac{1}{\mathbb{P}[T \geq t]} \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}[t \leq T < t + \Delta t]}{\Delta t}. \\ &= \frac{f(t)}{1 - F(t)}, \end{aligned}$$

where the second equality follows from Bayes' theorem. □

The *cumulative hazard function*, H, of hazard h is defined by

$$H(t) = \int_0^t h(s) \mathrm{d}s.$$

The hazard function h is of such importance in survival modeling because we can express the survival function (and thereby also the incidence function) in terms of the hazard. To see this, notice that $S'(t) = -f(t)$ and observe that

$$-\frac{\mathrm{d} \log(S(t))}{\mathrm{d}t} = \frac{f(t)}{S(t)} = h(t),$$

by Proposition 1.1. Taking integrals on both sides and rearranging yields the following useful expression of survival S:

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

1.2 Right-Censored Data

In practice, we may not observe the failure time T. This can happen because we may only observe the process until some (finite) point in time, and if the process has not failed by that time, we have no information on the time of its eventual failure. Suppose that the censoring (i.e. the point in time after which we stop to observe the process) takes place at some time C. The *observed* survival time (or time at risk), Y, is then defined by

$$Y = T \wedge C = \min\{T, C\},$$

which means that if the process fails before censoring time C, the failure time is observed via $Y = T$. Conversely, if the process has not yet failed at censoring time

C, the time at risk Y is equal to C and we have no information on the failure time T. We say that the observed time at risk Y is *right-censored*. Right-censoring can happen either by design (e.g. in a clinical trial which has a fixed ending time) or involuntarily due to losses to follow-up. Denote by $\delta = \mathbb{1}\{T \leq C\}$ an observed failure indicator which takes the value one if the process fails before the censoring time.

Since it is unrealistic to assume that we observe a process infinitely long until its eventual failure, survival model such as Cox proportional hazard models typically assume that the observed survival time is right-censored.

2 Ordinary Cox Proportional Hazard Modeling

2.1 Setup

A *Cox proportional hazard model* (Cox, 1972) attempts to explain the observed right-censored time at risk $Y = T \wedge C$ by some explanatory variables which are collected in a p-dimensional random vector $\mathbf{X}$. In proportional hazard modeling, we specify the hazard function h in (1) by using the semi-parametric specification

$$h(t; \mathbf{X}, \boldsymbol{\beta}) = h_0(t) \exp\left(\mathbf{X}^\top \boldsymbol{\beta}\right) \tag{2}$$

for fixed, but unknown coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$. The function h_0 is also a hazard function (i.e. it satisfies the right-hand side of (1)), but it is completely unspecified and does not depend on anything but time t. We refer to h_0 as the *baseline hazard*. Hence, there exists an unspecified *baseline cumulative hazard function* $H_0(t) = \int_0^t h_0(s)\mathrm{d}s$ and baseline survival $S_0(t) = \exp(-H_0(t))$.

With the proportional hazard specification in (2) the cumulative hazard of h satisfies

$$\begin{aligned} H(t; \mathbf{X}, \boldsymbol{\beta}) &= \int_0^t h_0(s) \exp\left(\mathbf{X}^\top \boldsymbol{\beta}\right) \mathrm{d}s \\ &= \exp\left(\mathbf{X}^\top \boldsymbol{\beta}\right) H_0(t), \end{aligned}$$

and the associated survival function is given by

$$\begin{aligned} S(t; \mathbf{X}, \boldsymbol{\beta}) &= \exp\Big(- \exp\left(\mathbf{X}^\top \boldsymbol{\beta}\right) H_0(t)\Big) \\ &= \Big(\exp\big(- H_0(t)\big)\Big)^{\exp(\mathbf{X}^\top \boldsymbol{\beta})} \\ &= S_0(t)^{\exp(\mathbf{X}^\top \boldsymbol{\beta})}. \end{aligned} \tag{3}$$

2.2 Fitting a Proportional Hazards Model

Suppose we observe a random sample $\big\{(\mathbf{X}_i, Y_i, \delta_i)\big\}_{i=1}^n$. Assume for now that for all failing individuals (i.e. individuals i for which $\delta_i = 1$), the failure times Y_i are unique. The goal is to estimate the unknown coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^p$ in the

proportional hazards specification (2). For this purpose, we consider the partial likelihood function

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{\{i\in[n]:\delta_i=1\}} \frac{h(Y_i;\mathbf{X}_i,\boldsymbol{\beta})}{\sum_{\{j\in[n]:Y_j\geq Y_i\}} h(Y_i|\mathbf{X}_j,\boldsymbol{\beta})} \\ &= \prod_{\{i\in[n]:\delta_i=1\}} \frac{\exp(\mathbf{X}_i^\top\boldsymbol{\beta})}{\sum_{\{j\in[n]:Y_j\geq Y_i\}} \exp(\mathbf{X}_j^\top\boldsymbol{\beta})}, \end{aligned} \tag{4}$$

which is called *partial* because it is computed only on individuals who have failed before the censoring time. Observe that the partial likelihood does *not* depend on the baseline hazard h_0, which substantially facilitates the optimization task. We maximize the partial likelihood by solving

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ -\frac{2}{n}\log L(\boldsymbol{\beta}) + \lambda_n P(\boldsymbol{\beta}) \right\}, \tag{5}$$

where P is an optional regularization penalty on the size of the coefficients, $\lambda_n \geq 0$ is a tuning parameter, and the scaling factor $2/n$ has been added for mathematical convenience. If the penalty P is convex, the optimization problem is convex, meaning that it can be solved easily. The value of tuning parameter λ_n can be determined via cross-validation. For P the elastic net penalty (Zou and Hastie, 2005), numerical details are described in Simon et al. (2011).

2.3 Estimating Survival

Suppose we have obtained an estimate $\widehat{\boldsymbol{\beta}}$ of the coefficient vector $\boldsymbol{\beta}$ in (2) by solving the optimization problem in (5). The goal is to estimate the survival function $S(t,\mathbf{X},\boldsymbol{\beta}) = S_0(t)^{\exp(\mathbf{X}^\top\boldsymbol{\beta})}$ in (3). For this purpose, we need, in addition to $\widehat{\boldsymbol{\beta}}$, an estimate of the unspecified baseline survival function S_0. Estimating S_0 is typically done non-parametrically, for instance by using a Nelson-Aalen or Kaplan-Meier estimator, which are calculated using $\{(Y_i,\delta_i)\}_{i=1}^n$ [Add reference]. With an estimate $\widehat{S}_0$ of S_0, we can estimate the survival function S via

$$\widehat{S}(t,\mathbf{X},\widehat{\boldsymbol{\beta}}) = \widehat{S}_0(t)^{\exp(\mathbf{X}^\top\widehat{\boldsymbol{\beta}})}.$$

2.4 What if the Times at Risk are not Unique?

Consider a situation where, for individuals for which $\delta_i = 1$, some of the times at risk Y_i are not unique. Breslow (1975) and Efron (1977) propose two different approaches for this situation. In the following, we briefly discuss the approach of Breslow (1975).

Let the sets $\mathcal{D}_i = \{j \in [n] : Y_j = Y_i\}$ contain the observations whose times at risk are tied with the one of individual i. Then, the likelihood function L in (4) becomes

$$L(\boldsymbol{\beta}) = \prod_{\{i\in[n]:\delta_i=1\}} \frac{\sum_{\{j\in\mathcal{D}_i\}} \exp(\mathbf{X}_j^\top\boldsymbol{\beta})}{\left(\sum_{\{j\in[n]:Y_j\geq Y_i\}} \exp(\mathbf{X}_j^\top\boldsymbol{\beta})\right)^{|\mathcal{D}_i|}}.$$

3 Competing Risk Modeling

3.1 Setup

There might be more than one failure type of a process. Models in which there are multiple causes of failure are referred to as *competing risk models*. Suppose that there are K failure types/causes in total and we observe a variable ε that indicates the failure type. Without loss of generality, assume that ε is supported on the set $\{1, \dots, K\}$ and $\varepsilon = k$ means that the process fails due to cause k. Assume further that we observe the right-censored time at risk $Y = T \wedge C$ and the product variable $\delta\varepsilon$. It holds that $\delta\varepsilon = k$ if the process fails before censoring time C due to cause k and that $\delta\varepsilon = 0$ if the process does not fail before censoring time C.

An important quantity in competing risk models are cumulative incidence functions. Unlike an ordinary incidence functions F, cumulative incidence functions consider each failure type separately. Hence, if there are K types of failure, there are K cumulative incidence functions, denoted F_k, for $k = 1, \dots, K$. Formally, for some time $t \in [0, \infty]$, the *cumulative incidence function of failure type* $k \in \{1, \dots, K\}$ is defined by

$$F_k(t) = \mathbb{P}[T \leq t, \varepsilon = k].$$

This definition[1] gives rise to the following decomposition of the overall incidence function F:

$$F(t) = \mathbb{P}[T \leq t] = \sum_{k=1}^{K} \mathbb{P}[T \leq t, \varepsilon = k] = \sum_{k=1}^{K} F_k(t).$$

Hence, for the survival function S it holds that

$$S(t) = 1 - F(t) = 1 - \sum_{k=1}^{K} F_k(t).$$

We emphasize that for survivors (for which $\delta\varepsilon = 0$), no cumulative incidence function is considered.

There are two main ways of specifying competing risk models, *proportional cause-specific hazard* models and *proportional subdistribution hazard* models.

3.2 Proportional Cause-Specific Hazard Models

3.2.1 Defining Cause-Specific Hazard

In proportional cause-specific hazard models, we essentially model all K failure types separately. Fix a failure type $k \in \{1, \dots, K\}$. The associated *cause-specific hazard function* h_k is defined by

$$h_k(t) = \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}[t \leq T < t + \Delta t, \varepsilon = k | T \geq t]}{\Delta t}. \tag{6}$$

[1] A cumulative incidence function F_k with $K > 1$ is not a distribution function, because $\lim_{t \to +\infty} F_k(t) \neq 1$. Hence, F_k does *not* have a density.

The cause-specific hazard $h_k(t)$ is interpreted as the instantaneous rate of failure due to type k, provided that the process has not yet failed.

Proposition 3.1. *Let random variable T denote a failure time and let F be its incidence function. Assume that there are K types of failure and let F_k be the cumulative incidence function of the k-th failure type. For the associated cause-specific hazard function h_k, it holds that*

$$h_k(t) = \frac{F_k'(t)}{1 - F(t)} = \frac{F_k'(t)}{S(t)}.$$

Proof. Analogous to the proof of Proposition 1.1 upon noticing that

$$F_k'(t) = \lim_{\Delta t \downarrow 0} \frac{F_k(t + \Delta t) - F_k(t)}{\Delta t} = \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}[t \leq T < t + \Delta t, \varepsilon = k]}{\Delta t}.$$

□

The associated *cause-specific cumulative hazard*, H_k, of h_k is then defined by

$$H_k(t) = \int_0^t h_k(s)\mathrm{d}s.$$

We can express the overall survival function in terms of the cause-specific hazard. Notice that $S'(t) = -\sum_{k=1}^K F'(t)$ and observe that

$$-\frac{\mathrm{d}\log(S(t))}{\mathrm{d}t} = \frac{\sum_{k=1}^K F_k'(t)}{S(t)} = \sum_{k=1}^K h_k(t),$$

by Proposition 3.1. Taking integrals on both sides and rearranging yields the following useful expression of overall survival S:

$$S(t) = \exp\left(-\sum_{k=1}^K \int_0^t h_k(s)\mathrm{d}s\right) = \exp\left(-\sum_{k=1}^K H_k(t)\right).$$

3.2.2 Specifying Cause-Specific Hazard

We specify the cause-specific hazard (6) by means of a Cox proportional hazard specification:

$$h\left(t; \mathbf{X}, \boldsymbol{\beta}^{(k)}\right) = h_{0,k}(t) \exp\left(\mathbf{X}^\top \boldsymbol{\beta}^{(k)}\right), \tag{7}$$

where $\boldsymbol{\beta}^{(k)} \in \mathbb{R}^p$ is a fixed, but unknown coefficient vector which is specific to failure type k. The function $h_{0,k}$ is also a cause-specific hazard function (i.e. it satisfies the right-hand side of (6)), but it is completely unspecified and does not depend on anything but time t. We refer to $h_{0,k}$ as the *cause-specific baseline hazard*. Hence, there exists an unspecified *cause-specific baseline cumulative hazard function* $H_{0,k}(t) = \int_0^t h_{0,k}(s)\mathrm{d}s$ and baseline cause-specific survival $S_{0,k}(t) = \exp(-H_{0,k}(t))$.

With the cause-specific proportional hazard specification in (7) the cumulative hazard of h_k satisfies

$$\begin{aligned} H_k\left(t; \mathbf{X}, \boldsymbol{\beta}^{(k)}\right) &= \int_0^t h_{0,k}(s) \exp\left(\mathbf{X}^\top \boldsymbol{\beta}^{(k)}\right) \mathrm{d}s \\ &= \exp\left(\mathbf{X}^\top \boldsymbol{\beta}^{(k)}\right) H_{0,k}(t). \end{aligned}$$

The overall survival function with the proportional hazards specification (7) depends on all K cause-specific coefficients and baseline survivals:

$$\begin{aligned} S\left(t; \mathbf{X}, \boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(K)}\right) &= \exp\left(-\sum_{k=1}^K H_k\left(t; \mathbf{X}, \boldsymbol{\beta}^{(k)}\right)\right) \\ &= \prod_{k=1}^K S_{0,k}(t)^{\exp\left(\mathbf{X}^\top \boldsymbol{\beta}^{(k)}\right)}. \end{aligned} \tag{8}$$

3.2.3 Fitting a Cause-Specific Proportional Hazard Model

Suppose we observe a random sample $\left\{(\mathbf{X}_i, Y_i, \delta_i \varepsilon_i, \delta_i)\right\}_{i=1}^n$. Assume for now that for all failing individuals (i.e. individuals i for which $\delta_i = 1$), the failure times Y_i are unique. For each failure type k, the goal is to estimate the unknown coefficient vector $\boldsymbol{\beta}^{(k)} \in \mathbb{R}^p$ in the cause-specific proportional hazard specification (7). For this purpose, we consider the cause-specific partial likelihood function

$$\begin{aligned} L_k\left(\boldsymbol{\beta}^{(k)}\right) &= \prod_{i=1}^n \left[\frac{h_k(Y_i; \mathbf{X}_i, \boldsymbol{\beta}^{(k)})}{\sum_{\{j \in [n]: Y_j \geq Y_i\}} h(Y_i | \mathbf{X}_j, \boldsymbol{\beta}^{(k)})}\right]^{\delta_i \mathbb{1}\{\varepsilon_i = k\}} \\ &= \prod_{i=1}^n \left[\frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta}^{(k)})}{\sum_{\{j \in [n]: Y_j \geq Y_i\}} \exp(\mathbf{X}_j^\top \boldsymbol{\beta}^{(k)})}\right]^{\delta_i \mathbb{1}\{\varepsilon_i = k\}}, \end{aligned}$$

which is maximized by solving

$$\widehat{\boldsymbol{\beta}}^{(k)} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ -\frac{2}{n} \log L_k(\boldsymbol{\beta}) + \lambda_n P(\boldsymbol{\beta}) \right\}. \tag{9}$$

Observe that all K problems (9) can be solved independently of each other and that solving (9) is equivalent to solving (5) by only using individuals for which $\delta_i \mathbb{1}\{\varepsilon_i = k\} = 1$. Thus, if the failure times of some failing individuals are not unique, one may instead use the adapted likelihood described in Section 2.4.

3.2.4 Estimating Survival

Suppose we have obtained estimates $\widehat{\boldsymbol{\beta}}^{(k)}$ of the coefficient vectors $\boldsymbol{\beta}^{(k)}$ in (7) by solving the optimization problem in (9) for each failure type k. The goal is to estimate the survival function in (8). For this purpose, we need, in addition to $\widehat{\boldsymbol{\beta}}^{(k)}$, an estimate of the unspecified case-specific baseline survival function $S_{0,k}$.

Again, $S_{0,k}$ can be estimated non-parametrically via e.g. a Kaplan-Meier estimator. [TODO: check input of nonparametric estimation of k-th survival function; paper isn't clear on this.] With estimates $\widehat{S}_{0,k}$ of $S_{0,k}$, we can estimate the survival function S via

$$\widehat{S}\left(t; \mathbf{X}, \widehat{\boldsymbol{\beta}}^{(1)}, \ldots, \widehat{\boldsymbol{\beta}}^{(K)}\right) = \prod_{k=1}^{K} \widehat{S}_{0,k}(t)^{\exp\left(\mathbf{X}^\top \widehat{\boldsymbol{\beta}}^{(k)}\right)}.$$

3.3 Proportional Subdistribution Hazard Models

3.3.1 Defining Subdistribution Hazard

This section builds on Fine and Gray (1999). Fix failure type $k \in \{1, \ldots, K\}$. The associated *subdistribution hazard function* h_k is defined by

$$h_k(t) = \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}\left[t \leq T \leq t + \Delta t, \varepsilon = k \mid (T \geq t) \cup ((T \leq t) \cap (\varepsilon \neq k))\right]}{\Delta t}. \tag{10}$$

The subdistribution hazard h_k is interpreted as the instantaneous rate of failure due to type k, provided that either the process has not yet failed *or* it has failed, but due to causes other than cause k.

Proposition 3.2. *Let random variable T denote a failure time and let F be its incidence function. Assume that there are K types of failure and let F_k be the cumulative incidence function of the k-th failure type. For the associated subdistribution hazard function h_k, it holds that*

$$h_k(t) = \frac{F_k'(t)}{1 - F_k'(t)} = -\frac{d \log(1 - F_k(t))}{dt}. \tag{11}$$

Proof. Follows immediately by Bayes' rule, the definition of the derivative $F_k'(t)$, and upon realizing that

$$\begin{aligned}\mathbb{P}\left[(T \geq t) \cup ((T \leq t) \cap (\varepsilon \neq k))\right] &= \mathbb{P}[T \geq t] + \mathbb{P}[(T \leq t) \cap (\varepsilon \neq k)] \\ &= 1 - F(t) + \sum_{j \neq k} F_j(t) = 1 - F_k(t).\end{aligned}$$

□

The associated *subdistribution cumulative hazard*, H_k, of h_k is then defined by

$$H_k(t) = \int_0^t h_k(s) \mathrm{d}s.$$

We can express the overall survival function in terms of the subdistribution hazard. Taking integrals in equation (11) and rearranging yields

$$F_k(t) = 1 - \exp\left(-H_k(t)\right).$$

Hence,

$$S(t) = 1 - \sum_{k=1}^{K} F_k(t) = 1 - K + \sum_{k=1}^{K} \exp\left(-H_k(t)\right).$$

3.3.2 Specifying Subdistribution Hazard

We specify the subdistribution (10) by means of a Cox proportional hazard specification:

$$h\left(t;\mathbf{X},\boldsymbol{\beta}^{(k)}\right)=h_{0,k}(t)\exp\left(\mathbf{X}^{\top}\boldsymbol{\beta}^{(k)}\right), \tag{12}$$

where $\boldsymbol{\beta}^{(k)}\in\mathbb{R}^p$ is a fixed, but unknown coefficient vector which is specific to failure type k. The function $h_{0,k}$ is also a subdistribution hazard function (i.e. it satisfies the right-hand side of (10)), but it is completely unspecified and does not depend on anything but time t. We refer to $h_{0,k}$ as the *subdistribution baseline hazard*. Hence, there exists an unspecified *subdistribution baseline cumulative hazard function* $H_{0,k}(t)=\int_0^t h_{0,k}(s)\mathrm{d}s$ and subdistribution baseline survival $S_{0,k}(t)=\exp(-H_{0,k}(t))$.

With the subdistribution proportional hazard specification in (12), the cumulative hazard of h_k satisfies

$$\begin{aligned}H_k\left(t;\mathbf{X},\boldsymbol{\beta}^{(k)}\right)&=\int_0^t h_{0,k}(s)\exp\left(\mathbf{X}^{\top}\boldsymbol{\beta}^{(k)}\right)\mathrm{d}s\\&=\exp\left(\mathbf{X}^{\top}\boldsymbol{\beta}^{(k)}\right)H_{0,k}(t).\end{aligned} \tag{13}$$

The overall survival function with the proportional hazards specification (12) depends on all K cause-specific coefficients and subdistribution baseline survivals:

$$\begin{aligned}S\left(t;\mathbf{X},\boldsymbol{\beta}^{(1)},\ldots,\boldsymbol{\beta}^{(K)}\right)&=1-K+\sum_{k=1}^{K}\exp\left(-H_k\left(t;\mathbf{X},\boldsymbol{\beta}^{(k)}\right)\right)\\&=1-K+\sum_{k=1}^{K}S_{0,k}(t)^{\exp\left(\mathbf{X}^{\top}\boldsymbol{\beta}^{(k)}\right)}.\end{aligned} \tag{14}$$

3.3.3 Fitting a Subdistribution Hazard Model

Suppose we observe a random sample $\left\{(\mathbf{X}_i,Y_i,\delta_i\varepsilon_i,\delta_i)\right\}_{i=1}^n$, where $Y_i=T_i\wedge C_i$ and censoring times C_i are observed. Assume for now that for all failing individuals (i.e. individuals i for which $\delta_i=1$), the failure times Y_i are unique. For each failure type k, the goal is to estimate the unknown coefficient vector $\boldsymbol{\beta}^{(k)}\in\mathbb{R}^p$ in the subdistribution proportional hazard specification (12).

We first need some additional definitions. Let

$$N_{i,k}(t)=\mathbb{1}\{T_i\le t,\varepsilon_i=k\}\quad\text{and}\quad Z_{i,k}(t)=1-N_{i,k}(t-).$$

Observe that for any Borel measurable function $g:[0,\infty)\to[0,\infty)$, it holds that $\int_0^t g(s)\mathrm{d}N_{i,k}(s)=g(T_i)N_{i,k}(t)$. Further,

- if $\varepsilon_i=k$, then $N_{i,k}(t)=\mathbb{1}\{T_i\le t\}$ and $Z_{i,k}(t)=\mathbb{1}\{T_i\ge t\}$;
- if $\varepsilon_i\neq k$, then $N_{i,k}(t)=0$ and $Z_{i,k}(t)=1$.

Define furthermore by

$$r_i(t) = \mathbb{1}\{C_i \geq (T_i \wedge t)\}$$

an indicator denoting knowledge of vital status of individual i at time t. It holds that

- if $\delta_i = 0$, then $r_i(t) = \mathbb{1}\{t \leq Y_i\}$;
- if $\delta_i = 1$, then $r_i(t) = 1$;
- if $r_i(t) = 1$, then $Z_{i,k}(t)$ and $N_{i,k}(t)$ are observed;
- if $r_i(t) = 0$, then $Z_{i,k}(t)$ and $N_{i,k}(t)$ are *not* observed;
- the products $r_i(t)Z_{i,k}(t)$ and $r_i(t)N_{i,k}(t)$ are always observed for $r_i(t) = 0, 1$.

Moreover, we define

$$w_i(t) = r_i(t)\frac{\widehat{G}(t)}{\widehat{G}(Y_i \wedge t)},$$

where $\widehat{G}$ is the Kaplan-Meier estimator of $G(t) = \mathbb{P}[C \geq t]$, that is, the survival function of censoring time C. The estimate $\widehat{G}$ is computed using $\{(Y_i, 1 - \delta_i)\}_{i=1}^n$.

Let the set

$$\mathcal{I}_k = \{i \in [n] : \delta_i \varepsilon_i = k\}$$

contain all individuals which we observe to fail due to cause k. For some $i \in \mathcal{I}_k$, let the set

$$\mathcal{R}_{i,k} = \{j \in [n] : (Y_j \geq Y_i) \text{ or } (0 < \delta_j \varepsilon_j \neq k)\}$$

contain all individuals who are at least as long at risk as individual i *or* fail due to causes other than cause k. We consider the pseudo likelihood of failure type k by

$$\begin{aligned} L_k\left(\boldsymbol{\beta}^{(k)}\right) &= \prod_{i=1}^{n}\left[\frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta}^{(k)})}{\sum_{j=1}^n Z_{j,k}(Y_i)w_j(Y_i)\exp(\mathbf{X}_j^\top \boldsymbol{\beta}^{(k)})}\right]^{\mathbb{1}\{\delta_i\varepsilon_i = k\}} \\ &= \prod_{i \in \mathcal{I}_k}\left[\frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta}^{(k)})}{\sum_{j \in \mathcal{R}_{i,k}} Z_{j,k}(Y_i)w_j(Y_i)\exp(\mathbf{X}_j^\top \boldsymbol{\beta}^{(k)})}\right]. \end{aligned} \tag{15}$$

Proposition 3.3. *The pseudo likelihood in (15) is observed.*

Proof. Note that we only need to compute the terms in the product in (15) for observations in the set $\mathcal{I}_k$. For fixed $i \in \mathcal{I}_k$ and $j \in [n]$, it holds for the individual summands in the denominator that

$$Z_{j,k}(Y_i)w_j(Y_i) = \begin{cases} \mathbb{1}\{Y_j \geq Y_i\}\frac{\widehat{G}(Y_i)}{\widehat{G}(Y_i \wedge Y_j)} & \text{if } \delta_j = 0, \\ \mathbb{1}\{Y_j \geq Y_i\}\frac{\widehat{G}(Y_i)}{\widehat{G}(Y_i \wedge Y_j)} & \text{if } \delta_j = 1, \varepsilon_j = k, \\ \frac{\widehat{G}(Y_i)}{\widehat{G}(Y_i \wedge Y_j)} & \text{if } \delta_j = 1, \varepsilon_j \neq k, \end{cases}$$

where all components are observed. Since the summation in the denominator of (15) is restricted to individuals in $\mathcal{R}_{i,k}$, we can further simplify the previous display. For $j \in \mathcal{R}_{i,k}$, it holds that

$$Z_{j,k}(Y_i)w_j(Y_i) = \begin{cases} 1 & \text{if } (Y_j \geq Y_i) \cap (\delta_j = 0), \\ 1 & \text{if } (Y_j \geq Y_i) \cap (\delta_j \varepsilon_j = k), \\ 1 & \text{if } (Y_j \geq Y_i) \cap (0 < \delta_j \varepsilon_j \neq k), \\ \widehat{G}(Y_i)/\widehat{G}(Y_j) & \text{if } (Y_j < Y_i) \cap (0 < \delta_j \varepsilon_j \neq k), \end{cases}$$

where all components are observed. □

The pseudo likelihood in (15) is maximized by solving

$$\widehat{\boldsymbol{\beta}}^{(k)} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ -\frac{2}{n} \log L_k(\boldsymbol{\beta}) + \lambda_n P(\boldsymbol{\beta}) \right\}. \tag{16}$$

Problem (16) is convex if the (optional) penalty P is convex. [TODO: If failure times are not unique, can we use adapted likelihood in Section 2.4?]

3.3.4 Estimating Survival

Suppose we have obtained estimates $\boldsymbol{\beta}^{(k)}$ of the coefficient vectors $\boldsymbol{\beta}^{(k)}$ in (12) by solving the optimization problem in (16) for each failure type k. The goal is to estimate the survival function in (14).

Fine and Gray (1999) propose to estimate the subdistribution baselinse cumulative hazard of failure type k, $H_{0,k}$, by using a Breslow (1975)-type estimator

$$\begin{aligned} \widehat{H}_{0,k}\left(t; \widehat{\boldsymbol{\beta}}^{(k)}\right) &= \sum_{i=1}^{n} \int_0^t \frac{w_i(u)}{\sum_{j=1}^n Z_{j,k}(u) w_j(u) \exp(\mathbf{X}_j^\top \widehat{\boldsymbol{\beta}}^{(k)})} \mathrm{d}N_{i,k}(u) \\ &= \sum_{i=1}^{n} \frac{w_i(T_i) N_{i,k}(t)}{\sum_{j=1}^n Z_{j,k}(T_i) w_j(T_i) \exp(\mathbf{X}_j^\top \widehat{\boldsymbol{\beta}}^{(k)})} \\ &= \sum_{i \in \mathcal{I}_k} \frac{w_i(T_i) N_{i,k}(t)}{\sum_{j \in \mathcal{R}_{i,k}} Z_{j,k}(T_i) w_j(T_i) \exp(\mathbf{X}_j^\top \widehat{\boldsymbol{\beta}}^{(k)})}. \end{aligned} \tag{17}$$

Proposition 3.4. *The estimator in (17) is observed for all times $t \in [0, \infty]$.*

Proof. We first observe that for any $i \in [n]$, it holds that

$$w_i(T_i)N_{i,k}(t) = \begin{cases} \mathbb{1}\{Y_i \leq t\} & \text{if } \delta_i = 1, \varepsilon_i = k, \\ 0 & \text{if } \delta_i = 0, \\ 0 & \text{if } \delta_i = 1, \varepsilon_i \neq k, \end{cases}$$

where all components are observed. Since the outer summation in (17) only sums over individuals in $\mathcal{I}_k$, the above reduces to $w_i(T_i)N_i(t) = \mathbb{1}\{Y_i \leq t\}$ for $i \in \mathcal{I}_k$.

Concerning the sum in the denominator in (17), for fixed $i \in \mathcal{I}_k$, it holds for any $j \in [n]$ that

$$Z_{j,k}(T_i)w_j(T_i) = \begin{cases} \mathbb{1}\{Y_j \geq Y_i\} & \text{if } \delta_j = 0, \\ \mathbb{1}\{Y_j \geq Y_i\} & \text{if } \delta_j = 1, \varepsilon_j = k, \\ \frac{\widehat{G}(Y_i)}{\widehat{G}(Y_i \wedge Y_j)} & \text{if } \delta_j = 1, \varepsilon_j \neq k, \end{cases}$$

where all components are observed. Since the sum in the denominator in (17) only sums over individuals in $\mathcal{R}_{i,k}$, we can simplify the previous display. For $j \in \mathcal{R}_{i,k}$, it holds that

$$Z_{j,k}(T_i)w_j(T_i) = \begin{cases} 1 & \text{if } (Y_j \geq Y_i) \cap (\delta_j = 0), \\ 1 & \text{if } (Y_j \geq Y_i) \cap (\delta_j \varepsilon_j = k), \\ 1 & \text{if } (Y_j \geq Y_i) \cap (0 < \delta_j \varepsilon_j \neq k), \\ \widehat{G}(Y_i)/\widehat{G}(Y_j) & \text{if } (Y_j < Y_i) \cap (0 < \delta_j \varepsilon_j \neq k), \end{cases}$$

where all components are observed. □

With the estimator in (17), we can estimate the subdistribution cumulative hazard function in (13) by

$$\begin{aligned} \widehat{H}_k\left(t; \mathbf{X}, \widehat{\boldsymbol{\beta}}^{(k)}\right) &= \int_0^t \exp\left(\mathbf{X}^\top \widehat{\boldsymbol{\beta}}^{(k)}\right) \mathrm{d}\widehat{H}_{0,k}\left(s; \widehat{\boldsymbol{\beta}}^{(k)}\right) \\ &= \exp\left(\mathbf{X}^\top \widehat{\boldsymbol{\beta}}^{(k)}\right) \widehat{H}_{0,k}\left(t; \widehat{\boldsymbol{\beta}}^{(k)}\right). \end{aligned}$$

It follows that the cumulative incidence function of failure type t can be estimated by $\widehat{F}_k\left(t; \mathbf{X}, \widehat{\boldsymbol{\beta}}^{(k)}\right) = 1 - \exp\left(-\widehat{H}_k\left(t; \mathbf{X}, \widehat{\boldsymbol{\beta}}^{(k)}\right)\right)$ and that overall survival (14) can be estimated by

$$\widehat{S}\left(t; \mathbf{X}, \widehat{\boldsymbol{\beta}}^{(1)}, \ldots, \widehat{\boldsymbol{\beta}}^{(K)}\right) = 1 - K + \sum_{k=1}^{K} \exp\left(-\widehat{H}_k\left(t; \mathbf{X}, \widehat{\boldsymbol{\beta}}^{(k)}\right)\right).$$

References

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review*, 43(1):45–57.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509.

Simon, N., Friedman, J. H., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, 39(5):1–13.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.