Survival Models

Max Welz
welz@ese.eur.nl

Econometric Institute
Erasmus School of Economics

September 10, 2021

1 Essential Theory

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. Let F be the distribution function of Y and f be its corresponding density. By definition, for some time $t \in [0, \infty]$,

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

measures the probability that the process fails before or at time t. Conversely, the survival function S, defined by

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

is the probability that failure occurs *after* time t. We call the distribution function F the *incidence function*. An essential quantity in survival modeling is the *hazard function* $h : \mathbb{R} \to \mathbb{R}$, defined by

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

The hazard function $h(t)$ is interpreted as the instantaneous rate of failure in individuals who are still at risk at time t. We emphasize that the hazard function is *not* a probability, which is a frequent misconception. For some time $t \in [0, \infty]$, the *cumulative hazard function* of hazard h is correspondingly defined by

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

By definition, we can relate the incidence function F to hazard h and survival S through the identity

$$F(t) = \int_0^t h(u) S(u) \mathrm{d}u.$$

Observe that the definition of the hazard function h gives rise to a differential equation of distribution F, namely $h(t) = \frac{F'(t)}{1-F(t)}$. Solving yields the following useful identity, which expresses distribution F in terms of cumulative hazard H:

$$F(t) = 1 - \exp\big(-H(t)\big).$$

Thus, we can also express survival S in terms of cumulative hazard H:

$$S(t) = \exp\big(-H(t)\big).$$

2 Cox Proportional Hazard Modeling

A *Cox proportional hazard model* (Cox (1972); hereafter Cox model) attempts to explain the survival time T by some explanatory variables which are collected in a p-dimensional random vector X. An essential component of a Cox model is the baseline hazard function h_0, which is a pre-specified hazard function that only depends on time instead of variables in X. The corresponding cumulative baseline hazard and baseline survival functions of h_0 are denoted by H_0 and S_0, respectively. Often, h_0 is chosen to be modeled non-parameterically, for instance via a Kaplan-Meier or a Nelson-Aalen estimator (see Cameron and Trivedi (2005) for definitions of these estimators).

With pre-specified baseline hazard and time $t \in [0, \infty]$, the hazard function of a Cox model is given by the semi-parameteric expression

$$h_\beta(t|X) = h_0(t)\exp\big(X^\top\beta\big), \tag{1}$$

where $\beta = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$ is a fixed but unknown vector of coefficients. Observe that we make the hazard's dependence on the vectors X and β explicit by expressing it as a conditional function of X and by using β as a subscript, respectively. By definition, it holds for the cumulative hazard H_β of h_β that

$$\begin{aligned} H_\beta(t|X) &= \int_0^t h_0(s)\exp\big(X^\top\beta\big)\mathrm{d}s \\ &= \exp\big(X^\top\beta\big)H_0(t) \end{aligned}$$

Thus, the associated survival function S_β of a Cox model is by definition given by

$$\begin{aligned} S_\beta(t) &= \exp\Big(-\exp\big(X^\top\beta\big)H_0(t)\Big) \\ &= \Big(\exp\big(-H_0(t)\big)\Big)^{\exp(X^\top\beta)} \\ &= S_0(t)^{\exp(X^\top\beta)}. \end{aligned} \tag{2}$$

3 Fitting a Cox Proportional Hazard Model

3.1 Setup

Suppose we have information on n observations, indexed $i = 1, \dots, n$. Suppose further that we observe non-negative real-valued random variables Y_i that measure the time at risk of individual i, as well as p-dimensional random vectors X_i which contain explanatory variables for the Y_i.

We assume that the times at risk Y_i are right-censored. This means that we assume the existence of latent variables T_i and C_i. The latent variable T_i denotes the failure time of individual i and the latent variable C_i denotes the censoring time of individual i. For the sake of simplicity, we assume that all individuals have the same censoring time, $C = C_i$, for all $i \in [n]$, and that C is observed; one may think of C as the ending time of a trial. Therefore, for the observed time Y_i, the identity $Y_i = T_i \wedge C$ holds. We say that individuals which have not yet failed at censoring time C are *survivors* (within observed time period), whereas we refer to individuals which fail before censoring time C as *failures*. Thus, for all survivors, it holds that $Y_i = C$, and for all failures, we have that $Y_i < C$. Thereupon, define a binary random variable δ_i that takes the value one if individual i is a failure and the value zero if it is a survivor, that is, $\delta_i = \mathbb{1}\{T_i < C\}$. The goal is to use the observed random sample $\big\{(X_i, Y_i, \delta_i)\big\}_{i=1}^n$ to estimate the unknown coefficient vector $\beta \in \mathbb{R}^p$ in the Cox hazard function in (1). We do so via maximum likelihood estimation.

Assume for the moment that all times at risk Y_i are unique; see Section 3.2 for a discussion on non-unique times at risk. To perform maximum likelihood estimation, we consider the partial likelihood function L, which is calculated on the failures and constructed as

$$\begin{aligned} L(\beta) &= \prod_{\{i\in[n]:\delta_i=1\}} \frac{h_\beta(Y_i|X_i)}{\sum_{\{j\in[n]:Y_j\geq Y_i\}} h_\beta(Y_i|X_j)} \\ &= \prod_{\{i\in[n]:\delta_i=1\}} \frac{\exp(X_i^\top \beta)}{\sum_{\{j\in[n]:Y_j\geq Y_i\}} \exp(X_j^\top \beta)}, \end{aligned} \tag{3}$$

with the hazard function h_β as in (1). Observe that the the partial likelihood does *not* depend on the baseline hazard h_0, as corresponding expressions cancel in the first line of the previous display.

We maximize (3) by maximizing its corresponding log-likelihood, minus a regularization penalty P_α, which depends on some pre-specified $\alpha \in [0, 1]$. The strength of the sparsity-enforcing penalty P_α is controlled via a fixed tuning parameter $\lambda_n \geq 0$. Thus, we obtain estimator $\widehat{\beta}$ of β in (1) by solving

$$\widehat{\beta} = \arg\max_{\beta\in\mathbb{R}^p} \left\{ \frac{2}{n} \left[\sum_{\{i\in[n]:\delta_i=1\}} \left(X_i^\top \beta - \ln \left(\sum_{\{j\in[n]:Y_j\geq Y_i\}} \exp\left(X_j^\top \beta\right) \right) \right) \right] - \lambda_n P_\alpha(\beta) \right\}, \tag{4}$$

where the scaling factor $2/n$ has been added for mathematical convenience, and P_α is the elastic net penalty (Zou and Hastie, 2005), defined by

$$P_\alpha(\beta) = \alpha \sum_{j=1}^p |\beta_j| + \frac{1}{2}(1-\alpha) \sum_{j=1}^p \beta_j^2 = \alpha \|\beta\|_1 + \frac{1}{2}(1-\alpha)\|\beta\|_2^2.$$

The problem in (4) is convex for all choices of $\alpha \in [0, 1]$ and $\lambda_n \geq 0$, hence it can be solved easily. The value of tuning parameter λ_n can be determined via cross-validation. Numerical details are described in Simon et al. (2011).

With estimate $\widehat{\beta}$, we can estimate the Cox model's survival function in (2) by

$$\widehat{S}(t) = S_{\widehat{\beta}}(t) = \widehat{S}_0(t)^{\exp\left(X^\top \widehat{\beta}\right)}.$$

Recall that the baseline survival S_0 does not depend on β, hence its estimate $\widehat{S}_0$ also does not depend on $\widehat{\beta}$. Thus, as previously discussed, $\widehat{S}_0$ is typically estimated separately in non-parameteric fashion. An estimate of the Cox model's hazard function in (1) can be constructed analogously.

TODO: Add some stuff on the proportional hazard assumption and merge with main document

3.2 What if the Times at Risk are not Unique?

Consider a situation where 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 (3) becomes

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

and the optimization problem in (4) is adapted correspondingly.

4 Competing Risk Modeling

There might be several causes for an individual to fail. Suppose that there are K failure types/causes in total and that there exist variables ε_i that indicate the cause of failure of individual i. Without loss of generality, assume that ε_i have support on the set $\{1, \ldots, K\}$ and $\varepsilon_i = k$ means that individual i fails due to cause k. In practice, we observe the variables $\delta_i \varepsilon_i$. Hence, if individual i survives, we observe $\delta_i \varepsilon_i = 0$, whereas if individual i fails before the censoring time, we observe $\delta_i \varepsilon_i = \varepsilon_i$. Obviously, $\delta_i \varepsilon_i$ is supported on $\{0, 1, \ldots, K\}$. Models in which there are multiple causes of failure are referred to as *competing risk models*. In such models, we observe the random sample $\{(X_i, Y_i, \delta_i, \delta_i \varepsilon_i)\}_{i=1}^n$.

In situations with competing risk, one is typically only interested in one single cause of failure. For instance, in a trial on cardiovascular diseases, one is typically only interested in cardiovascular deaths and not in non-cardiovascular deaths (some individuals in the trial might die of causes other than cardiovascular diseases). A naive approach in such a situation is to artificially set the time at risk of all non-cardiovascular deaths equal to the censoring time, thereby effectively counting them as survivors. However, this approach may produce upward biased estimates of incidence function F, which will in turn lead to downwards biased estimates of the survival function S (e.g. Austin et al., 2016).

Approaches that provides accurate estimates of incidence and survival despite the presence of competing risks typically make use of *Cumulative Incidence Functions*. Unlike 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 $k = 1, \ldots, K$. Mathematically, for some time $t \in [0, \infty]$, the *cumulative incidence function of failure type* $k \in \{1, \ldots, K\}$ is defined by

$$F_k(t) = \mathbb{P}[T_1 \leq t, \varepsilon_1 = k].$$

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

$$F(t) = \mathbb{P}[T_1 \leq t] = \sum_{k=1}^{K} \mathbb{P}[T_1 \leq t, \varepsilon_1 = k] = \sum_{k=1}^{K} F_k(t).$$

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 $\delta_i \varepsilon_i = 0$), no cumulative incidence function is considered.

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

4.1 Proportional Cause-Specific Hazard Models

Proportional cause-specific hazard models essentially fit K separate Cox proportional hazard models. Hence, for cause $k \in [K]$, the hazard function associated with this cause is given by

$$\begin{aligned} h^k_{\beta^k}(t|X_i) &= \lim_{\Delta \downarrow 0} \frac{\mathbb{P}[t \leq T_i < t + \Delta, \varepsilon_i = k | T_i \geq t]}{\Delta} \\ &= h^k_0(t) \exp(X^\top \beta^k), \end{aligned}$$

where h^k_0 is a Breslow-type of the baseline hazard of individuals for which $\varepsilon_i = k$. We use the k-superscript in $\beta^k \in \mathbb{R}^p$ to remind us that the coefficients of each of the K. Consequently, we estimate β^k by solving the problem (4) on individuals for which $\varepsilon_i = k$. We can subsequently estimate the associated survival function $S(t) = S_{(\beta^1, \ldots, \beta^K)}(t)$ by

$$\hat{S}(t) = S_{(\hat{\beta}^1, \ldots, \hat{\beta}^K)}(t) = \prod_{k=1}^{K} S^k_0(t)^{\exp(X^\top \widehat{\beta}_k)}$$

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

make dependence on X explicit: we need info on t and X for prediction! We need to estimate beta K times

References

Austin, P. C., Lee, D. S., and Fine, J. P. (2016). Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609.

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

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge University Press.

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.

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.