Survival Models

Max Welz
welz@ese.eur.nl

Econometric Institute
Erasmus School of Economics

September 8, 2021

1 Essential Theory

Suppose we want to model the failure rate of a some statistical process. Concretely, let the random variable Y 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 $t \in [0, \infty]$,

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

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. A popular choice for F is the exponential distribution. An essential quantity in survival modeling is the *hazard function* $h : \mathbb{R} \to \mathbb{R}$, defined by

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

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 defined by

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

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 Y 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-parametric 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 failure time of observation i, as well as p-dimensional random vectors X_i which contain explanatory variables for the Y_i. Assume that the failure times Y_i are right-censored at a fixed time $T > 0$, meaning that observations which have not yet failed at censoring time T are counted as *survivors*. Thus, for all survivors, it holds that $Y_i = T$, and for all failures, we have that $Y_i < T$. 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. The goal is to use the 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 failure times Y_i are unique; see Section 3.2 for a discussion on non-unique failure times. 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 Failure Times are not Unique?

Consider a situation where some of the failure times 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 failure times 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.

References

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.