Survival analysis and censored data *

- we observe either the survival time T or else the censoring time $C \Rightarrow Y = \min(T, C)$
- we also observe a status indicator $\delta = \begin{cases} 1 & \text{if } T < C \\ 0 & \text{if } T \geq C \end{cases}$ then if $\delta = 1$ we observe the true survival time, if $\delta = 0$, we observe the censoring time.

we hence have n (Y, δ) pairs $(Y_1, \delta_1) \dots (Y_n, \delta_n)$

In general, we need to assume that the censoring mechanism is independent: conditional on the features, the event time T is independent of the censoring time C. we often have to consider the data collection in order to determine whether independent censoring is a reasonable assumption.

The Kaplan-Meier Survival curve

the survival curve (or function) is defined as $S(t) = P_r(T > t)$.

we let $d_1 < d_2 < \dots < d_k$ denote the k unique death times among the non censored patients and we let q_k denote the number of patients who died at time d_k. For $k = 1, \dots, k$ we let n_k denote the numbers of patients alive and in the study just before d_k; these are the at risk patients. The set of patients that are at risk at a given time are referred to as the risk set

$$P_r(T > d_k) = P_r(T > d_k \mid T > d_{k-1}) P_r(T > d_{k-1}) + \underbrace{P_r(T > d_k \mid T < d_{k-1})}_{0} P_r(T \leq d_{k-1})$$

$$S(d_k) = P_r(T > d_k) = P_r(T > d_k \mid T > d_{k-1}) \times \dots \times P_r(T > d_2 \mid T > d_1) P_r(T > d_1) \quad \text{with } \hat{P}_r(T > d_j \mid T > d_{j-1}) = \frac{n_j - q_j}{n_j}$$

$$\hat{S}_k = \prod_{j=1}^{k} \left(\frac{n_j - q_j}{n_j} \right)$$ and for times t between d_k and d_{k+1}, we set $\hat{S}(t) = \hat{S}(d_k)$

The log-rank test

We examine how the events in each group unfold sequentially in time. In order to test H_0: $E(X) = \mu$ for some random variable X, one approach is to construct a test statistic of the form $W = \frac{X - \mu}{\sqrt{\text{Var}(X)}}$

	group 1	group 2	Total
died	q_{1k}	q_{2k}	q_k
survived	$n_{1k} - q_{1k}$	$n_{2k} - q_{2k}$	$n_k - q_k$
total	n_{1k}	n_{2k}	n_k

$X = \sum_{k=1}^{K} q_{1k}$ is $\mu = \sum_{k=1}^{K} \frac{n_{1k}}{n_k} q_k$

$$\Rightarrow W = \frac{\sum_{k=1}^{K} \left(q_{1k} - \frac{q_k}{n_k} n_{1k} \right)}{\sqrt{\sum_{k=1}^{K} \frac{q_k (n_{1k}/n_k)(1 - n_{1k}/n_k)(n_k - q_k)}{n_{k-1}}}}$$

when the sample size is large, W has a standard normal distribution: this can be used to compute a p-value for the null hypothesis that there is no difference between the survival curves in the two groups.

Regression models with survival curves

The hazard function (or rate) is defined as $h(t) = \lim_{\Delta t \to 0} \frac{P_r(t \leq T \leq t + \Delta t \mid T > t)}{\Delta t}$ where T is the (unobserved) survival time. $h(t) = \lim_{\Delta t \to 0} \frac{P_r((t < T \leq t + \Delta t) \cap (T > t)) / \Delta t}{P_r(T > t)} = \frac{P_r(t < T \leq t + \Delta t)/\Delta t}{P_r(T > t)} = \frac{f(t)}{S(t)}$

where $f(t) = \lim_{\Delta t \to 0} \frac{P_r(t < T \leq t + \Delta t)}{\Delta t}$ is the probability density function associated with T: i.e the instantaneous rate of death at time t.

The likelihood associated with the i^{th} observation is

$L_i \begin{cases} f(y_i) & \text{if the } i^{th} \text{ observation is not censored} \\ S(y_i) & \text{if the } i^{th} \text{ observation is censored} \end{cases}$

$= f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}$ The intuition is: if $Y = y_i$ and i^{th} observation is not censored then the likelihood is the probability of dying in a time interval around time y_i. If the i^{th} observation is censored, then the likelihood is the probability of surviving at least until time y_i.

To model the survival time as function of the covariates $\Rightarrow h(t|x_i) = \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})$

$\Rightarrow h(t|x_i) = h_0(t) \exp(\sum_{j=1}^{p} x_{ij}\beta_j)$ $\Rightarrow$ one unit increase in x_{ij} corresponds to an increase in $h(t|x_i)$ by a factor of $\exp(\beta_j)$

- The partial likelihood $PL(\beta) = \prod_{i:\delta_i=1} \frac{\exp(\sum_{j=1}^{p} x_{ij}\beta_j)}{\sum_{i':y_{i'} \geq y_i} \exp(\sum_{j=1}^{p} x_{i'j}\beta_j)}$. To estimate β, we simply maximize PL with respect to β.

Shrinkage for the Cox model

$-\log\left(\prod_{i:\delta_i=1} \frac{\exp(\sum_{j=1}^{p} x_{ij}\beta_j)}{\sum_{i':y_{i'} \geq y_i} \exp(\sum_{j=1}^{p} x_{i'j}\beta_j)}\right) + \lambda P(\beta)$ is minimized with respect to $\beta = (\beta_1, \ldots \beta_p)^T$,

if $P(\beta) = \sum_{j=1}^{p} \beta_j^2 =$ ridge, $\sum_{j=1}^{p} |\beta_j| =$ lasso.