Survival Analysis and Censored Data

By

AINDRILA GARAI

MSC STATISTICS, IIT KANPUR

aindrilag22@iitk.ac.in

Survival and Censoring Times:

The **survival time**(T) represents the time at which the event of interest occurs- for instance, the time at which the patient dies or the customer cancels his or her subscription.

The **censoring time**(C) is the time at which censoring occurs- for example, the time at which the patient drops out of the study or the study ends.

We do work with $Y = \min(T,C)$ and an indicator variable

$$\delta = \begin{cases} 1 & \text{if } T \leq C \quad \text{(left-censoring)} \\ 0 & \text{if } T > C \quad \text{(right-censoring)} \end{cases}$$

- Suppose an example, An analysis that does not take into consideration the reason why the patients dropped out will likely overestimate the true average survival time. - Similarly, suppose that males who are very sick are more likely to drop out of the study than females who are very sick. Then a comparison of male and female survival times may wrongly suggest that males survive longer than females.

In general, let us assume the event time T is independent of the censoring time C but it is difficult to determine from the data.

We focus more on right-censoring.

The Kaplan-Meier Survival Curve:

The survival curve or survival function is defined as- $S(t) = \Pr(T > t)$ (decreasing function) which quantifies the probability of surviving past time t.

Eg. Suppose that a company is interested in modeling customer churn. Let T represent the time that a customer cancels a subscription to the company's service. Then $S(t)$ represents the probability that a customer cancels later than time t. **The larger the value of $S(t)$, the less likely that the customer will cancel before time t.**

- $S(t)$ is complicated by the presence of censoring. So, we let $d_1 < d_2 < \cdots < d_K$ denote the K unique death times among the noncensored patients, and we let q_k denote the number of patients who died at time d_k. For $k = 1, \ldots, K$, we let r_k denote the

number of patients alive and in the study just before d_k; these are the at risk patients referred to as the risk set. By calculation we get,

$$S(d_k) = \Pr(T > d_k \mid T > d_{k-1}) \times \cdots \times \Pr(T > d_2 \mid T > d_1)\Pr(T > d_1)$$

- The Kaplan-Meier estimator of the survival curve is

$$\widehat{\Pr}(T > d_j \mid T > d_{j-1}) = (r_j - q_j)/r_j$$

and then

$$\hat{S}(d_k) = \prod_{j=1}^{k}\left(\frac{r_j - q_j}{r_j}\right)$$

- For times t between d_k and d_{k+1}, we set $\hat{S}(t) = \hat{S}(d_k)$.
- The Kaplan-Meier survival curve has a step-like shape.

The Log-Rank Test:

To compute the log-rank test statistic, we calculate-

$$W = \frac{\sum_{k=1}^{K}(q_{1k} - \mathrm{E}(q_{1k}))}{\sqrt{\sum_{k=1}^{K}\mathrm{Var}\,(q_{1k})}} = \frac{\sum_{k=1}^{K}\left(q_{1k} - \frac{q_k}{r_k}r_{1k}\right)}{\sqrt{\sum_{k=1}^{K}\frac{q_k(r_{1k}/r_k)(1 - r_{1k}/r_k)(r_k - q_k)}{r_k - 1}}}.$$

where q_{1k} is number of patients of group 1 who died and r_{1k} is the risk set of group 1.

- When the sample size is large, the log-rank test statistic W has approximately a standard normal distribution.
- Computing p- value we can conclude our decision.

Regression Models With a Survival Response:

- The hazard function(rate of death) or hazard rate also known as the force of mortality is formally defined as-

$$h(t) = \lim_{\Delta t \to 0}\frac{\Pr(t < T \le t + \Delta t \mid T > t)}{\Delta t}$$

where T is the (unobserved) survival time.

- The relationship between the hazard function $h(t)$ the survival function $S(t)$ is

$$h(t) = \frac{f(t)}{S(t)}$$

where $f(t) = \lim_{\Delta t \to 0}\frac{\Pr(t<T\le t+\Delta t)}{\Delta t}$ and $S(t) = \Pr(T > t)$.

- The likelihood associated with the ith observation is

$$L_i = \begin{cases} f(y_i) & \text{if the } i \text{ th observation is not censored} \\ S(y_i) & \text{if the } i \text{ th observation is censored} \end{cases}$$
$$= f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}$$

  Assuming $f(t)$ is exponential or Gamma or Weibell distribution, we can find the pararmeter by MLE.

- One possible approach is to assume a functional form for the hazard function $h(t|xi)$, such as $h(t \mid x_i) = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)$ where the exponent function guarantees that the hazard function is non-negative. We then calculate $S(t|xi)$ and the parameter by MLE.

Proportional Hazards:

The proportional hazards assumption states that

$$h(t \mid x_i) = h_0(t)\exp\left(\sum_{j=1}^{p} x_{ij}\,\beta_j\right)$$

where $h_0(t) \geq 0$ is an unspecified function, known as the baseline hazard. The exponential term is called relative risk for the feature vector $x_i = \left(x_{i1}, \ldots, x_{ip}\right)^T$.

Baseline hazard function:

We make no assumptions about its functional form. The hazard function is very flexible and can model a wide range of relationships between the covariates and survival time. Our only assumption is that a one-unit increase in x_{ij} corresponds to an increase in $h(t|xi)$ by a factor of $\exp(\beta_j)$.

Cox's Proportional Hazards Model:

The magic of Cox's proportional hazards model lies in the fact that it is in fact possible to estimate β without having to specify the form of $h_0(t)$.

the total hazard at time yi for the at risk observations is

$$\sum_{i':y_{i'} \geq y_i} h_0\,(y_i)\exp\left(\sum_{j=1}^{p} x_{i'j}\,\beta_j\right)$$

The partial likelihood is simply the product of these probabilities over all of the uncensored observations-

$$PL(\beta) = \prod_{i:\delta_i=1} \frac{h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{ij}\,\beta_j\right)}{\sum_{i':y_{i'} \geq y_i} h_0\,(y_i)\exp\left(\sum_{j=1}^{p} x_{i'j}\,\beta_j\right)}$$

Remark:

- In the case of a single binary covariate Cox's proportional hazards model is exactly equal to the log-rank test.
- An intercept can be absorbed into the baseline hazard $h_0(t)$.
- We have assumed that there are no tied failure times. In the case of ties, the exact form of the partial likelihood is a bit more complicated and a number of computational approximations must be used.
- We can estimate the survival curve $S(t|x)$ for an individual with feature vector x by estimating the baseline hazard $h_0(t)$.

Shrinkage for the Cox Model:

Here, we consider minimizing a penalized version of the negative log partial likelihood

$$-\log\left(\prod_{i:\delta_i=1}\frac{\exp\left(\sum_{j=1}^{p}x_{ij}\,\beta_j\right)}{\sum_{i':y_{i'}\geq y_i}\exp\left(\sum_{j=1}^{p}x_{i'j}\,\beta_j\right)}\right)+\lambda P(\beta)$$

wrt $\beta=\left(\beta_1,\ldots,\beta_p\right)^T$ where $P(\beta)=\sum_{j=1}^{p}\beta_j^2$ or $P(\beta)=\sum_{j=1}^{p}|\beta_j|$

i) when $\lambda>0$, then minimizing thr above eq. yields a shrunken version of the coefficient estimates.

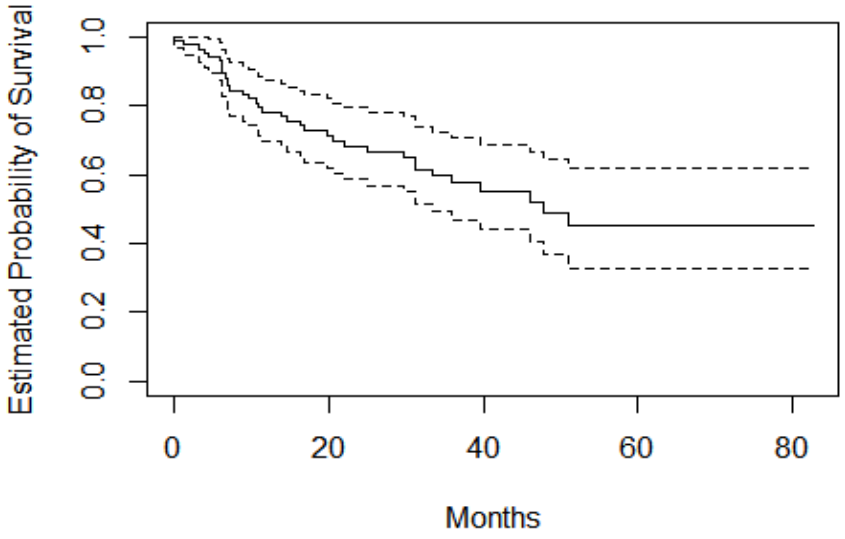
ii) For a sufficiently large value of λ, using a lasso penalty will give some coefficients that are exactly equal to zero.

- To assess the model fit which involves stratifying the observations using the coefficient estimates. In particular, for each test observation, we compute the "risk" score We then use these risk scores to categorize the observations based on their "risk".

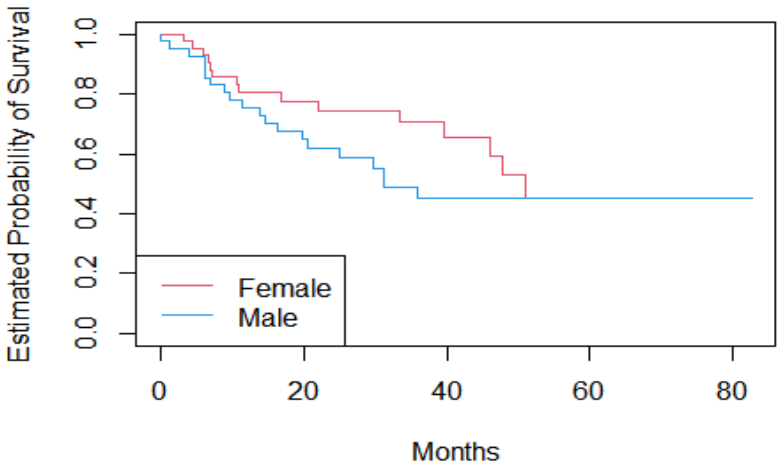
<u>Now try to plot:</u>

```
library (ISLR2)
attach (BrainCancer)
table (sex)

## sex
## Female   Male
##     45     43

table (diagnosis)
```

```
## diagnosis
## Meningioma  LG glioma  HG glioma      Other
##         42          9         22         14

table (status)

## status
##  0  1
## 53 35

library (survival)
fit.surv <- survfit ( Surv (time, status) ~ 1)
plot (fit.surv , xlab = " Months ", ylab = " Estimated Probability of Surviva
l ") # Kaplan-Meier survival curve
```



```
fit.sex <- survfit ( Surv (time, status) ~ sex)
plot (fit.sex , xlab = " Months ", ylab = " Estimated Probability of Survival
", col = c(2,4))
legend ("bottomleft", levels (sex), col = c(2,4), lty = 1)
```




```
logrank.test <- survdiff ( Surv (time, status) ~ sex) #  log-rank test
logrank.test
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ sex)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=Female 45       15     18.5     0.676      1.44
## sex=Male   43       20     16.5     0.761      1.44
##
##  Chisq= 1.4  on 1 degrees of freedom, p= 0.2
```

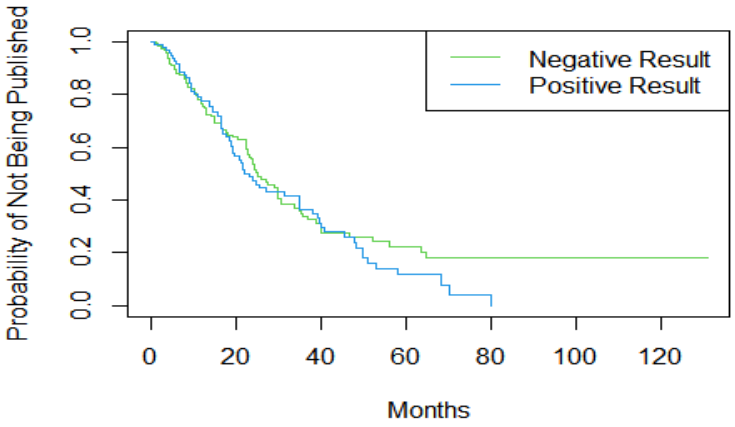
```
fit.cox <- coxph ( Surv (time, status) ~ sex)
summary (fit.cox)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ sex)
##
##   n= 88, number of events= 35
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## sexMale 0.4077    1.5033   0.3420 1.192    0.233
##
##         exp(coef) exp(-coef) lower .95 upper .95
## sexMale     1.503     0.6652     0.769     2.939
##
```

```
## Concordance= 0.565  (se = 0.045 )
## Likelihood ratio test= 1.44  on 1 df,   p=0.2
## Wald test            = 1.42  on 1 df,   p=0.2
## Score (logrank) test = 1.44  on 1 df,   p=0.2
```

Some work on Publication Data:

```
fit.posres <- survfit ( Surv (time, status) ~ posres , data = Publication )
plot (fit.posres , xlab = " Months ", ylab = " Probability of Not Being Publi
shed ", col = 3:4)
legend ("topright", c(" Negative Result ", " Positive Result ") ,
col = 3:4, lty = 1)
```




```
fit.pub <- coxph ( Surv (time, status) ~ posres , data = Publication)
fit.pub

## Call:
## coxph(formula = Surv(time, status) ~ posres, data = Publication)
##
##          coef exp(coef) se(coef)     z    p
## posres 0.1481    1.1596   0.1616 0.916 0.36
##
## Likelihood ratio test=0.83  on 1 df, p=0.3611
## n= 244, number of events= 156

logrank.test <- survdiff ( Surv (time, status) ~ posres , data = Publication)
logrank.test

## Call:
## survdiff(formula = Surv(time, status) ~ posres, data = Publication)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## posres=0 146       87     92.6     0.341     0.844
## posres=1  98       69     63.4     0.498     0.844
```

```
##
##  Chisq= 0.8  on 1 degrees of freedom, p= 0.4

fit.pub2 <- coxph ( Surv (time, status) ~ . - mech , data = Publication)
fit.pub2

## Call:
## coxph(formula = Surv(time, status) ~ . - mech, data = Publication)
##
##                coef  exp(coef)   se(coef)      z       p
## posres    5.708e-01  1.770e+00  1.760e-01  3.244 0.00118
## multi    -4.086e-02  9.600e-01  2.512e-01 -0.163 0.87079
## clinend   5.462e-01  1.727e+00  2.620e-01  2.085 0.03710
## sampsize  4.678e-06  1.000e+00  1.472e-05  0.318 0.75070
## budget    4.385e-03  1.004e+00  2.465e-03  1.779 0.07518
## impact    5.832e-02  1.060e+00  6.676e-03  8.735 < 2e-16
##
## Likelihood ratio test=149.2  on 6 df, p=< 2.2e-16
## n= 244, number of events= 156
```