Survival Analysis

Stat 526

April 13, 2018

1 Functions of Survival Time

Let T be the survival time for a subject. Then $P[T < 0] = 0$ and T is a continuous random variable. The *Survival function* is defined as

$$S(t) = P[T > t] = 1 - F(t).$$

It is clear that $S(0) = 1$ and $S(\infty) = 0$. The survival function can be explained by the probability of the subject to live longer than t.

The *hazard function* is defined by

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T \geq t)}{\Delta t}.$$

It is clear that

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

where f and F are density or distribution of T. It can be explained as an instant hazard of the people exposed.

The *cumulative hazard function* is defined as

$$H(t) = \int_0^t h(x)dx = -\log[S(t)].$$

It can be explained as the cumulative hazard of the people exposed until time t.

2 Some Well Known Distributions

The density function of the *exponential* distribution with parameter $\lambda > 0$ is

$$f(t) = \lambda e^{-\lambda t},$$

for $t > 0$. It is clear that $S(t) = e^{-\lambda t}$, $h(t) = \lambda$ and $H(t) = \lambda t$. This distribution is the constant hazard distribution.

The density of the *Weibull* distribution with parameter α and β is

$$f(t) = \alpha\gamma(\gamma t)^{\alpha-1} e^{-(\gamma t)^{\alpha}}.$$

It is cleat that when $\alpha = 1$ it is exponential distribution. The distribution is

$$F(t) = 1 - e^{-(\gamma t)^\alpha}.$$

It is clear that

$$S(t) = e^{-(\gamma t)^\alpha},$$

$$h(t) = \alpha\gamma e^{-(\gamma t)^\alpha} = \alpha\gamma(\gamma t)^{\alpha-1},$$

and

$$H(t) = (\gamma t)^\alpha.$$

Thus, for Weibull distribution

$$\log[-\log S(t)] = \alpha(\log\gamma + \log t).$$

Therefore, it is suggested to take a look at the estimated $\log[-\log S(t)]$ versus $\log t$ to diagnose Weibull distribution.

R defines the output for the parameters as

1. The scale parameter α, and
$$-\log(\gamma) = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_j.$$
2. If α are equal for all groups, then Weibull distribution has the proportional hazard property. In the default, **R** requires the scale parameters are the same.

If $\log(Y)$ follow a normal distribution, then we call Y follows a *log-normal* distribution. Therefore, the distribution function of log-normal with parameter μ and σ^2 is

$$F(t) = \Phi(\frac{\log t - \mu}{\sigma}).$$

The density function of *Gamma*-distribution with parameter α and β is

$$f(t) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t}.$$

3 Non-parametric Estimate the Survival Function

There is no assumption for the distributions. Let us look at the leukemia dataset In Table 1. There are total 42 observations in this dataset. The objective is to test the drug effect. Thus, 21 of them took drugs and 21 of them took placebo. The assignment for drugs or placebo are completely randomized.

The response is the time (weeks) that the patient live. The predictor censor tells us the corresponding patient drop off from the experiment at the time. Thus, we only know that the patient live longer than the time we record. In this dataset, 1 means not dropoff or relapse and 0 means dropoff or censored . In this dataset, no patients in placebo dataset dropped off.

Table 1: Survival Data For Leukemia

treat	censor	weeks	treat	censor	weeks
drug	0	6	placebo	1	1
drug	1	6	placebo	1	1
drug	1	6	placebo	1	2
drug	1	6	placebo	1	2
drug	1	7	placebo	1	3
drug	0	9	placebo	1	4
drug	0	10	placebo	1	4
drug	1	10	placebo	1	5
drug	0	11	placebo	1	5
drug	1	13	placebo	1	8
drug	1	16	placebo	1	8
drug	0	17	placebo	1	8
drug	0	19	placebo	1	8
drug	0	20	placebo	1	11
drug	1	22	placebo	1	11
drug	1	23	placebo	1	12
drug	0	25	placebo	1	12
drug	0	32	placebo	1	15
drug	0	32	placebo	1	17
drug	0	34	placebo	1	22
drug	0	35	placebo	1	23

Suppose there is no dropoff. Then, the empirical-survival function is

$$\hat{S}(t) = 1 - \frac{\#\{Patient > t\}}{\#Patient}$$

where $\#\{Patient > t\}$ means the number of the patient live longer than t.

Next, let us consider the dataset with censoring subjects. Let (T_i, C_i) be the record for observation i with observation (t_i, δ_i), where T_i is the time and C_i is the indicator of censoring. Assume that $t_1 \leq t_2 \leq \cdots \leq t_n$. Then, if observation i relapse $(\delta_i = 1)$ then the instant density is $f(t_i)$ and if observation i censors $(\delta_i = 0)$ then the observed probability is $S(t_i)$. Thus, we have the likelihood function is

$$L(t_1, \cdots, t_n) = \prod_{\delta_i=1} f(t_i) \prod_{\delta_i=0} S(t_i) = \prod_{\delta_i=1} h(t_i)^{\delta_i} \prod_{i=1}^{n} S(t_i).$$

Thus, we can assume

$$\hat{S}(t_i) = \prod_{j=1}^{i} (1 - \lambda_i),$$

where $0 \leq \lambda_1, \cdots, \lambda_n < 1$, because $S(t)$ is right-continuous. Then, the mass function at $t_1, \cdots, t_n$ is

$$\hat{f}(t_i) = \hat{S}(t_{i-1}) - \hat{S}_i = \lambda_i \prod_{j=1}^{i-1} (1 - \lambda_j).$$

Let d_i die and n_i happen at t_i. Then, the likelihood at t_i part is

$$\hat{L}_i = \hat{f}(t_i)^{d_i} \hat{S}(t_i)^{n_i - d_i} = \lambda_i^{d_i} (1 - \lambda_i)^{n_i - d_i}.$$

Thus, we have

$$L = \prod_i \lambda_i^{d_i} (1 - \lambda_i)^{n_i - d_i}.$$

Thus,

$$\hat{\lambda}_i = \frac{d_i}{n_i}.$$

Therefore, we have

$$\hat{S}(t) = \prod_{t_i \leq t} (1 - \frac{d_i}{n_i}),$$

where d_i is the number of death and n_i is the number of at risk (remaining). It is clear that

$$\log \hat{S}(t) = \sum_{t_i \leq t} \log(1 - \frac{d_i}{n_i}),$$

Thus,

$$\hat{Var}[\log \hat{S}(t)] = \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}.$$

There is another options (Greenwood's Formula)

$$\hat{Var}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}.$$

This estimator is called **Kaplan Meier** estimator. It is the ML estimator of the survival function. Let us take a look at the estimator of the treatment group for the leukemia dataset. The event happened at points: $6, 7, 9, 10, 11, 13, 16, 17, 19, 20, 22, 23, 25, 32, 34, 35$ totally 16 points. The number of death and at risk at the above time is $(3, 21), (1, 17), (0, 16), \cdots$. It is clear that when $t < 6$, $\hat{S}(t) = 1$; when $6 \leq t < 7$, $d_i = 3$ and $n_i = 21$,

$$\hat{S}(t) = (1 - \frac{3}{21}) = \frac{6}{7};$$

when $7 \leq t < 9$,

$$\hat{S}(t) = \frac{6}{7} \times (1 - \frac{1}{17});$$

and so on.

The significance test is based on a *log-rank* test. See detail in my output file. The analysis result for this data set tells us a significant drug effect exist because the p-value is 0.0000417.

4 Parametric Method

Let us still consider the leukemia dataset. If we assume the distribution is exponential, then there are two parameter λ_t and λ_p for treatment and placebo groups. Let us look at the treatment group. We have $h(t) = \lambda_t$ and $S(t) = e^{-\lambda_t t}$. Then, the likelihood function is

$$L = \lambda_t^{\sum \delta_i} \exp[-\lambda_t \sum_{i=1}^{n} t_i].$$

Thus,

$$\hat{\lambda}_t = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}.$$

For the treatment group, we have

$$\hat{\lambda}_t = \frac{9}{359}.$$

In R, the estimate of treatment is defined by

$$\hat{\beta}_t = -\log(\hat{\lambda}_t) = 3.69.$$

Similarly, we have $\hat{\beta}_p = 2.16$. They are significant different because the p-value is 0.000049.

Let us look at the Weibull distribution. The hazard function is

$$h(t) = \alpha \lambda^{\alpha} t^{\alpha - 1}.$$

It is clear that the hazard is a decreasing function of t if $\alpha < 1$ and a increasing function of t if $\alpha > 1$. When $\alpha > 1$, we call it is an *accelerated life* models.

It is not easy to compute the estimate of the parameters for Weibull distribution. Thus, a numerical method is suggested. R has a function to fit the Weibull model. R rewrite the model into

$$h(t) = \alpha t^{\alpha - 1} \exp[-\beta' x],$$

and call α scale parameter. The output tells us that $\hat{\alpha} = 0.7322$, and $\hat{\beta}_t = 3.5157$, $\hat{\beta}_p = 2.248$. Thus, the hazard of treatment group is less than the hazard of placebo group.

We assume the scale parameter α is the same for the two groups. Thus, if we rewrite $\alpha\beta'$ by $\tilde{\beta}$. Then, the hazard function becomes

$$h(t) = \alpha t^{\alpha-1} \exp(-\tilde{\beta}'x) = h_0(t) \exp(-\tilde{\beta}'x).$$

It is clear that $h_0(t)$ does not dependent on the covariates or treatment. We call this model proportional hazard model.

5 Semi-parametric method: proportional hazards model

The assumption for cox-proportional hazards model is

$$h(t) = h_0(t) e^{\beta' x},$$

where x is a vector of predictors and β is a vector of parameters independent of time t, h_0 is only a function of t; that is

$$\beta' x = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_j,$$

where x_j are covariates or treatment.

Cox suggests to look at the partial likelihood

$$L_p = \prod_{i=1}^{n} \frac{e^{\beta' x}}{\sum_{T \geq t_i} e^{\beta' x}}$$

to estimate the parameters. This function is called partial likelihood function. It is the hazards of observation at t_i over the sum of the hazard of all the remaining.

We can prove that the definition above is equivalent to

$$S(t) = [S_0(t)]^{e^{\beta' x}}.$$

The estimates are the values to maximize the partial likelihood function. In the leukemia example, $\beta_t = 0$ and $\beta_p = 1.57$. The p-value is 0.000014. Thus, the treatment is significant. Once the parameters are estimated, we can estimate the baseline hazard function. R has those functions.

In R output, we can directly read the survival function $\hat{S}(t)$ which should be a step function. Then, the hazard function at time t_i is

$$h(t_i) = \frac{S_{t_{i-1}} - S_{t_i}}{S_{t_i}};$$

and

$$\hat{H}(t_i) = \sum_{j \leq i} h(t_i)(t_{j+1} - t_j);$$

or

$$\hat{H}(t_i) = \sum_{j \leq i} h(t_i)(t_j - t_{j-1}).$$

The non-parametric Kaplan-Meier estimator is the MLE of survival function. It is suggested to compare the result with this estimator to diagnose the model.