Survival Analysis

Minato Nakazawa, Ph.D. (Professor, Kobe Univ. Grad. School Heallth Sci.)

July 20, 2022

Fully explaining survival analysis is difficult within this introductory class, but many functions to conduct survival analysis are provided by `survival` package.

1 Concept of survival analysis

In longitudinal observation of the effects by toxic substances, not only the changes of quantitative indices but also the time to event such as death could be used to evaluate the strength of the toxicity of that substances. The data like time to event can be analyzed by survival analysis (a.k.a. event history analysis).

Amongst one of the most famous methods is the Kaplan-Meier's product-limit estimate, which is the products of (1 − number of events divided by population at risk) at all times of occurences of events. The time when this value goes across 0.5 is median survival time. For the data of time to events for 2 groups, the logrank test or generalized Wilcoxon test can be used to test the difference between the 2 groups. Besides those nonparametric methods, there are parametric approaches to fit the time to events with any known distribution, like exponential distribution or Welbull distribution. The famous semi-parametric approach of survival analysis is the Cox's regression (a.k.a. fitting a proportional hazard model), which assume that the ith individual's hazard can be expressed as the product of the baseline hazard and $\exp(\sum \beta z_i)$, where z_i is the ith value of covariate vector z and β is coefficient's vector.

After typing `require(survival)` or `library(survival)`, `Surv()` generates the survival time object, `survfit()` will calculate the Kaplan-Meier's product-limit estimate (this result can also be used to draw survival curves), `survdiff()` will conduct the logrank test, and `coxph()` fit the proportional hazard model to the data. If you want to know the survival analysis in detail, you should read another text like Bull *et al.*, 1997.

2 Kaplan-Meier method

Let the times of event happening since beginning of time at risk $t_1, t_2, ...$, the numbers of events at each time $d_1, d_2, ...$, and the size of population at risk just before the each time $n_1, n_2, ...$ The size of population at risk decreases not only by the event occurrence but also by censoring such as moving out or loss to follow up or death by competing risks. When the censoring and event occurred at the same time, usually the censoring occurred just after the event happening.

Here the Kaplan-Meier's product-limit estimates $\hat{S}(t)$ can be defined as follows.

$$\hat{S}(t) = (1 - d_1/n_1)(1 - d_2/n_2)... = \prod_{i<t}(1 - d_i/n_i)$$

Clearly this value means the probability of survival and is numerically 1 at first (nobody has experienced the event) and 0 in the end (after the everybody experienced the event).

The standard error of $\hat{S}(t)$ is given by the Greenwood's formula shown below.

$$\mathrm{var}(\hat{S}) = \hat{S}^2 \times \sum_{i<t} \frac{d_i}{n_i(n_i - d_i)}$$

Estimated $\hat{S}(t)$ is usually plotted as survival curve with 95% confidence intervals[*1].

In Rgui console, basic grammer for Kaplan-Meier method is shown below (with comments). After loading survival package in memory by typing `library(survival)` or `require(survival)`, `dat <- Surv(times, flags)` generates the survival time data `dat`, where the flags for censoring become 1 when the observation ends by event's occurrence and become 0 when the observation ends by censoring. Conducting Kaplan-Meier method is simply `res <- survfit(dat~1)`, or `res <- survfit(dat~group)` if you want to estimate this by group. By typing `plot(res)`, we can get the graph of survival curves. Detailed result of estimation can be obtained by `summary(res)`.

practice

The `leukemia` data.frame in the `survival` package is the result of randomized controlled trial for the maintenance chemotherapy's effect to delay the remission of acute myelogenous leukemia. Included variables are the following 3.

time time to remission or censoring in weeks.

status flag for censoring, where 0 means censoring, 1 means remission.

x whether maintenance chemotherapy was conducted or not (`Maintained` or `Nonmaintained`).

Let's conduct Kaplan-Meier's method to estimate survival curves for the 2 groups (under maintenance chemotherapy or not). How long the median survival times (here, median times to remission) of the 2 groups are?

In Rgui console, type as follows.

```
require(survival)
print(res <- survfit(Surv(time, status)~x, data=leukemia))
plot(res, xlab="(Weeks)", lty=1:2, main="Periods until remission of acute myelogenous leukemia")
legend("right", lty=1:2, legend=levels(leukemia$x))
```

The second line conducts Kaplan-Meier estimation and shows result below.

```
                records n.max n.start events median 0.95LCL 0.95UCL
x=Maintained         11    11      11      7     31      18      NA
x=Nonmaintained      12    12      12     11     23       8      NA
```

The number of maintained patients is 11 and non-maintained patients is 12. Among them, remission cases were 7 and 11, respectively. Median periods until remission were 31 weeks for maintained patients and 23 weeks for non-maintained patients. Lower limits of 95% confidence intervals of the median periods until remission were 18 weeks and 8 weeks, respectively. Upper limits of those were both infinity. The third line draws survival curves as solid line and dashed line for maintained group and non-maintained group, respectively. The fourth line adds a legend to the graph.

In EZR, read the `leukemia` data set in `survival` package (Caution on 20th July 2022: `aml` and `leukemia` are included in `survival`, but those are not identified as data.frame, so that you have to need to execute `aml2 <- as.data.frame(aml)` after loading `survival` package). Select [File], [Read data set from an attached packages], then double-click `survival` in the top-left box, subsequently double-click `leukemia` in the top-right box. After that, click [OK].

Kaplan-Meier estimate will be done by select [Statistical analysis], [Survival analysis], [Kaplan-Meier survival curve and logrank test], then select [time] as "Time-to-event variable", select [status] as "Status indicator". As the "Grouping variable", [x] should be selected if you would like to do Kaplan-Meier estimate separately for maintained group and non-maintained group, otherwise leave there unselected.

This menu can draw the survival curve with confidence intervals by checking the box on. After setting all options, click [OK].

[*1] However, if $\hat{S}(t)$s were estimated for 2 groups and drawing both to compare them, confidence intervals are not usually drawn.

3 Logrank test

Here I will give a brief explanation about the concept of logrank test. Let's imagine 8 rats and randomly assign 2 groups (administering toxic substance A and B) to them, then follow them up. On the day 4, 6, 8, 9, the rats in the first group (which took toxic substance A) died. On the day 5, 7, 12, 14, the rats in the second group (which took toxic substance B) died. There is no censoring.

The concept of logrank test is, making 2 by 2 contingency tables of group and alive/dead at each time of event, and calculate the common chi-square value in a Cochran-Mantel-Haentzel's manner.

In the example shown above, let's denote expected number of death for jth group at the ith time of events as e_{ij}, observed total number of death at time i as d_i, population at risk of jth group at time i as n_{ij}, total population at risk at time i as n_i, then

$$e_{ij} = d_i \cdots n_{ij}/n_i$$

In the above example, $e_{11} = 1 \cdots 4/8 = 0.5$. Next, denote the number of death of jth group at time i as d_{ij}, the weight of time i as w_i, the score of jth group at time i as u_{ij}, then

$$u_{ij} = w_i \cdots (d_{ij} - e_{ij})$$

In the logrank test, every weights are 1. Then

$$u_{ij} = d_{ij} - e_{ij}$$

The summary score for group j, u_j can be calculated as follows.

$$u_j = \sum_i d_{ij} - e_{ij}$$

The variance V of the score can be considered as follows.

$$V = V_{jj} = \sum_i \frac{(n_i - n_{ij})n_{ij} \cdots d_i(n_i - d_i)}{{n_i}^2(n_i - 1)}$$

The chi-square value ${\chi_0}^2$, which obeys chi-square distribution of degree of freesom 1, is given by the formula below.

$${\chi_0}^2 = {u_1}^2/V$$

In the above example, $u_1 = (1-4/8)+(0-3/7)+(1-3/6)+(0-2/5)+(1-2/4)+(1-1/3)+(0-0/2)+(0-0/1) = 1.338...$ and $V = 1.568...$, then ${\chi_0}^2 = 1.338^2/1.568 = 1.14$. Because 1.14 is much smaller than 3.84, which is 95% point of chi-square distribution with degree of freedom 1, we cannot judge that there is significant difference between the two groups.

In Rgui console, the script conducting this is very simple as follows.

```
require(survival)
time <- c(4,6,8,9,5,7,12,14)
event <- rep(1,8)
group <- c(1,1,1,1,2,2,2,2)
survdiff(Surv(time,event) ~ group)
```

In the case of `leukemia` dataset in `survival` package, the script will be as follows.

```
require(survival)
survdiff(Surv(time, status) ~ x, data=leukemia)
```

The resulting p value is 0.0653, which means not significant difference between the maintained group and non-maintained group at the significance level of 5%.

In EZR, logrank test of the null-hypothesis that remission time are not different between maintained and nonmaintained groups is simultaneously conducted with the Kaplan-Meier estimate as already explained. Note: you can calculate the generalized Wilcoxon test in a manner of Peto-Peto instead of logrank test, by setting the radio-button to do so.

4 Cox regression

The Kaplan-Meier estimate and logrank test assume no specific distribution in the population, thus nonparametric method. Cox regression assumes that the individuals' hazards are "proportional" to the common baseline hazard. In that sense, it's semi-parametric method.

The basic idea of the Cox regression is: Denote the covariates' vectors affecting the occurrence of events as $z_i = (z_{i1}, z_{i2}, ..., z_{ip})$ for ith individual. Denote the instantaneous rate of the event occurrence for this individual at time t as $h(z_i, t)$. This is called as "hazard function". Cox regression assumes the following formula.

$$h(z_i, t) = h_0(t) \cdot \exp(\beta_1 z_{i1} + \beta_2 z_{i2} + ... + \beta_p z_{ip})$$

where $h_0(t)$ is the common baseline hazard, which is the instantaneous rate of event occurrence at time t of "base individual", who has no effect on event occurrence by all covariates. Unknown parameters $\beta_1, \beta_2, ..., \beta_p$ should be estimated. The effect of covariates on event occurrence is the proportional coefficients as $\exp(\beta_x z_{ix})$. This is called as "proportional hazard".

The original model by Cox considered the time-dependent covariates where z_i changes by time. However, usually we assume the effect of covariates on the event occurrence is independent from time (thus not changing by time). Therefore, the ratio of hazards between different individuals is constant regardless with time: The ratio of the 1st individual's hazard at time t to the 2nd individual's hazard at time t is not including $h_0(t)$ (cancelled from numerator and denominator), then the hazard ratio is given by the following formula. It means that the hazard ratio doesn't depend on the shape of $h_0(t)$.

$$\frac{\exp(\beta_1 z_{11} + \beta_2 z_{12} + ... + \beta_p z_{1p})}{\exp(\beta_1 z_{21} + \beta_2 z_{22} + ... + \beta_p z_{2p})}$$

The relationship of survival function and hazard function can be summarized as follows. Denote the non-negative random variable showing the time to event occurrence as T, then the survival function $S(t)$ is the probability of $T \geq t$. By this definition, $S(0) = 1$. The hazard function $h(t)$ is the instantaneous probability of event occurrence at time t. Then we can get the following equations.

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)} \\ &= -\frac{dS(t)}{dt} \frac{1}{S(t)} = -\frac{d(\log(S(t))}{dt} \end{aligned}$$

The cumulative hazard function $H(t)$ is,

$$H(t) = \int_0^t h(u)du = -\log S(t)$$

Thus we obtain

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

Denote the cumulative hazard function at time t of an individual with covariates vector z as $H(z,t)$, the survival function of the same individual as $S(z,t)$. If the proportional hazard stands,

$$H(z,t) = \int_0^t h(z,u)du = \int_0^t h_0(u)\exp(\beta z)du = \exp(\beta z)H_0(t)$$

Then we obtain the following equation.

$$S(z,t) = \exp(-H(z,t)) = \exp\{-\exp(\beta z)H_0(t)\}$$

Take the logarithm and inverse the sign and take the logarithm again, the we obtain the following equation.

$$\log(-\log S(z,t)) = \beta z + \log H_0(t)$$

From this equation, we can see that the pararell curves with the gap of βz will be drawn, being survival time as horizontal axis and $\log(-\log S(z,t))$ as vertical axis. If this pararell nature is not met, the "proportional hazard" assumption cannot stand and thus Cox regression is not suitable.

In estimation of β, the concept of partial likelihood is applied. If we decompose the probability of event occurring for ith individual at time t into the probability of single event occurrence at time t and the conditional probability that the event occurred for the specific individual i under the condition that the event occurred at time t, the former is still unknown unless we can assume any specific parametric model, but the latter $L(i,t)$ can be always estimated as the ratio of ith individual's hazard being numerator and the sum of all individuals' hazard within the whole population at risk at time t.

For the all event occurrences, denote the products of all $L(i,t)$s as L, the meaning of L is the total likelihood minus the likelihood concerning time, thus is called as partial likelihood. To estimate a "good" parameter β that asymptotically converges to the true parameter as the sample size becomes larger, and whose distribution obeys a normal distribution, and whose variance becomes smallest, Cox conjectured that such β could be obtained when the L became maximum and this conjecture was given prove by the Martingale theory. By this fact, the proportional hazard model is also known as Cox regression[*2]. The basic form of Cox regression in R is `coxph(Surv(time,cens)~grp+covar,data=dat)`.

Practice

In the `leukemia` dataset, conduct Cox regression on the effect of treatment (maintained / nonmaintained) on the survival time.

```
require(survival)
summary(res <- coxph(Surv(time,status)~x, data=leukemia))
KM <- survfit(Surv(time,status)~x, data=leukemia)
par(family="sans",las=1,mfrow=c(1,3))
plot(KM, lty=1:2, main="Kaplan-Meier plot of survival time of leukemia dataset.")
legend("topright", lty=1:2, legend=levels(leukemia$x))
plot(survfit(res),
 main="The survival curve of the reference individual\n with the treatments being covariates")
plot(KM, fun=function(y) {log(-log(y))}, lty=1:2, main="Double logarithmic plot of leukemia dataset")
```

The result given in the second line is shown below.

[*2] If multiple events occur simultaneously, there are several methods to treat them: Exact method, Efron's method, Breslow's method, discrete method, and so on. However, whenever possible, Exact method is recommended. The discrete method should be used when the survival times are given as discrete measures. Many statistical software uses Breslow's method, but the default method in R's `coxph()` function is Efron's method. Generally speaking, Efron's method gives closer results than Breslow's method.

```
Call:
coxph(formula = Surv(time, status) ~ x, data = leukemia)

  n= 23
                coef exp(coef) se(coef)    z     p
xNonmaintained 0.916       2.5    0.512 1.79 0.074

               exp(coef) exp(-coef) lower .95 upper .95
xNonmaintained       2.5        0.4     0.916      6.81

Rsquare= 0.137   (max possible= 0.976 )
Likelihood ratio test= 3.38  on 1 df,   p=0.0658
Wald test            = 3.2  on 1 df,   p=0.0737
Score (logrank) test = 3.42  on 1 df,   p=0.0645
```

The p value of the test of null-hypothesis that the maintained and nonmaintained group have the same hazard is 0.074, so that the null-hypothesis is not rejected at the 5% significance level*3. The `exp(coef)` 2.5 is the estimated hazard ratio of 2 groups, so that we can judge the nonmaintained group has 2.5 times higher hazard of maintained group's hazard but the 95% confidence intervals includes 1.

By the third line and later scripts, 3 graphs are drawn. From left to right, the Kaplan-Meier plot estimated for 2 groups separately, the baseline survival curve with 95% confidence intervals as the result of Cox regression with treatment being covariates, and the double logarithmic plot are drawn, respectively.

If you dare to draw the baseline survival curves of Cox regression with covariates for the 2 treatment groups separately, for example, `subset=(x=="Maintained")` option can be used in the `coxph()` function, when the group variable cannot be included as covariates. More than 2 survival curves could be drawn without erasing previous graphs by specifying `par(new=TRUE)` option. However, I don't recommend this manner.

There are three strategies to control the effect of covariates on the survival time. For example, to analyze the survival time of cancer patients, the effects of stage should be controlled. The possible strategies to control them are:

1. Analyzing the survival time separately by each stage.
2. Assuming that the effects of other covariates are common to all stages, then set the stage as strata.
3. Including the stage as covariates in the same model.

The third strategy has an advantage that the effect of stage can be quantitatively estimated, but it requires the unrealistic condition that the baseline hazards are the same for all stages. In addition, the method of coding stages as covariates may affect the result (usually coding as dummy variable).

The second strategy means that the baseline hazards are different by stage. In the `coxph()` function, the option `strata()` can be used for different baseline hazards. For example, if the data.frame of survival time of cancer patients is `leukemia`, which includes 4 variables: the variable of survival time `time`, the variable of censoring flag `status`, the group variable showing the treatment `x`, the variable of stage of cancer progress `stage`, then the model can be written as `coxph(Surv(time,status)~x+strata(stage), data=leukemia)`*4

Anyway, Cox regression is a kind of model-fitting, so that we can select better models by the residual analysis, likelihood ratio test, and the squared multiple correlation coefficients, but AICs are usually not available because calculating AIC requires the specified distribution in the baseline hazard.

*3 `Score (logrank) test` in the bottom line is the result of Rao's score test, different from the logrank test with `survdiff()`.

*4 However, in fact, `leukemia` data.frame does not include stage.

In EZR, select [Statistical analysis], [Survival analysis], [Cox proportional hazard regression], then select `time` in the box of "Time", select `status` in the box of "Event", and select `x` in the box of "Explanatory variables" (if you set here variables as the form of "+strata(`Variable Names`)", different baseline hazards by strata are assumed). After all, click [OK], then you get the result.