NONPARAPMETRIC DATA ANALYSIS

PH-REGRESSION AND SURVIVAL FUNCTION

Wenchen Guo

Contents

1 DATA DISCRIPTION

We'll use the NCTG Lung Cancer Data in the survival R package. The NCTG Lung Cancer Data is the records of survival in patients with advanced lung cancer from the North Central Cancer Treatment Group.

	inst	time	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
1	3.00	306.00	2.00	74.00	1.00	1.00	90.00	100.00	1175.00	
2	3.00	455.00	2.00	68.00	1.00	0.00	90.00	90.00	1225.00	15.00
3	3.00	1010.00	1.00	56.00	1.00	0.00	90.00	90.00		15.00
4	5.00	210.00	2.00	57.00	1.00	1.00	90.00	60.00	1150.00	11.00
5	1.00	883.00	2.00	60.00	1.00	0.00	100.00	90.00		0.00
6	12.00	1022.00	1.00	74.00	1.00	1.00	50.00	80.00	513.00	0.00
7	7.00	310.00	2.00	68.00	2.00	2.00	70.00	60.00	384.00	10.00
8	11.00	361.00	2.00	71.00	2.00	2.00	60.00	80.00	538.00	1.00
9	1.00	218.00	2.00	53.00	1.00	1.00	70.00	80.00	825.00	16.00
10	7.00	166.00	2.00	61.00	1.00	2.00	70.00	70.00	271.00	34.00

Table 1: Lung Cancer Data

The dataset contains 228 observations. Tables 1 shows the first 10 rows of the data.
There are 10 variables:

- inst: Institution code
- time: Survival time in days
- status: censoring status 1=censored, 2=dead
- age: Age in years
- sex: Male=1 Female=2
- ph.ecog: ECOG performance score (0=good 5=dead)
- ph.karno: Karnofsky performance score (bad=0-good=100) rated by physician
- pat.karno: Karnofsky performance score as rated by patient
- meal.cal: Calories consumed at meals
- wt.loss: Weight loss in last six months

(Performance scores rate how well the patient can perform usual daily activities.)

In this report, we only use variables time, status, and sex for analyzing.

2 Explore PH-Regression and Survival Function

2.1 Survival Function with Kaplan-Meier Estimator

2.1.1 Simple Survival Curve

First, we can use the survival function to create a simple survival curve that doesn't consider any different groupings.

```
Call: survfit(formula = surv ~ 1, data = lung, conf.int = 0.9)


     n events median 0.9LCL 0.9UCL
   228    165    310    285    353
```

The median survival time for the 228 observations is 310 days.

Plot the survival curve: Figure 1 is the survival curve. The horizontal axis (x-axis) represents time in days, and the vertical axis (y-axis) shows the probability of surviving or the proportion of people surviving. The solid line represents survival curve of the 228 observations. The two dash lines represent the lower and upper confidence limits for the curve.

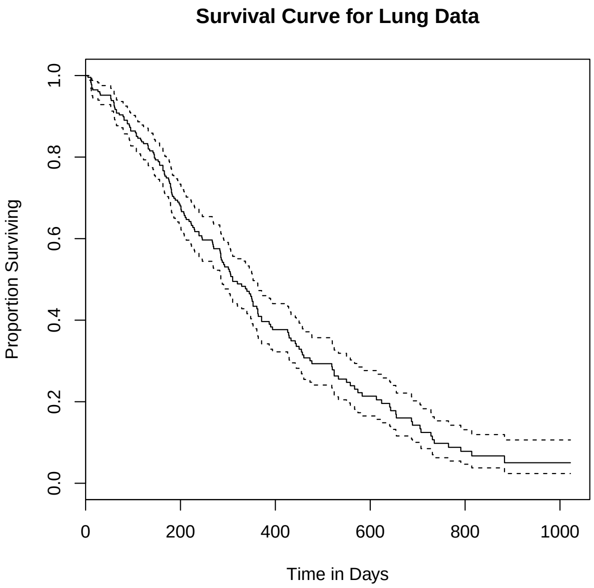

Figure 1: Survival Curve for Lung Data

2.1.2 Survival Curves Between Sex

Kaplan-Meier curves are good for visualizing differences in survival between two categorical groups. We use the survival function to create survival curves separately by sex. 1 stands for male and 2 stands for female.

```
Call: survfit(formula = surv ~ sex, data = lung, conf.int = 0.9)

        n events median 0.9LCL 0.9UCL
sex=1 138    112    270    222    306
sex=2  90     53    426    350    524
```

There are 138 males and 90 females in the data set. The results shows that the median survival time for female is 426 days, but for male is 270 days. Thus, it's reasonable to doult that males tend to have worse survival than females.

We can use the survival curve for further checking.
Plot the survival curve:

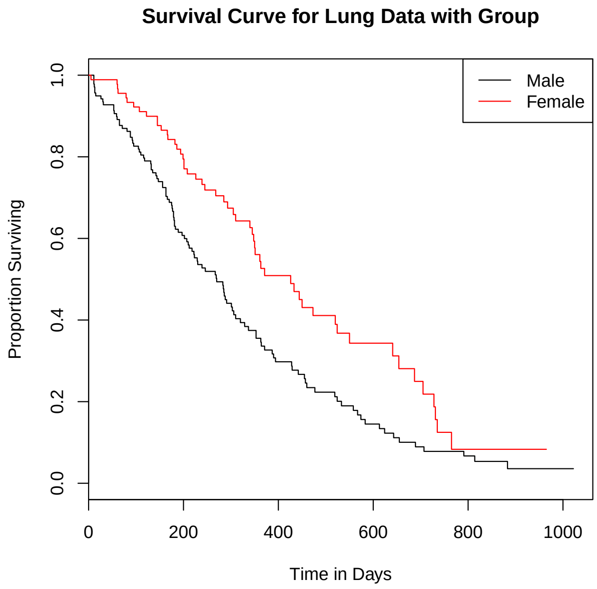

Figure 2: Survival Curve for Lung Data

Figure 2 is the survival curve with group. The black line represents survival curve of male observations. The red line represents survival curve of female observations. It is obvious that females have a better chance of surviving than males.

2.2 PH-Regression with Hazard-rate

Now we can use PH regression to estimate the reduction in hazard from male (baseline) to female. The Cox model (1972) is expressed by hazard function denoted by h(t), which can interpretate the risk of dying at certain time t. It can be estimate as follow:

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p)$$

where

- Term t indicates survival time.
- $\lambda(t)$ is hazard function based on a set of covariates $(x_1, x_2, \ldots, x_p)$ is (in this case, we have one covariate "sex" with p=1).
- Coefficients $(b_1, b_2, \ldots, b_p)$ measures the effect of covariates, and when x_i all equals to 0, the term $\lambda(t)$ is called "baseline hazard function".
- Notice that the partial likelihood does not depend on the underlying hazard function λ since $\frac{\lambda(t|Z=1)}{\lambda(t|Z=0)} = e^{\beta}$.

Cox analyses can be computed as follows:

```
Call:
coxph(formula = Surv(time, status) ~ sex, data = lung)

  n= 228, number of events= 165

       coef exp(coef) se(coef)      z Pr(>|z|)
sex -0.5310    0.5880   0.1672 -3.176  0.00149 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    exp(coef) exp(-coef) lower .90 upper .90
sex     0.588      1.701    0.4466    0.7741

Concordance= 0.579  (se = 0.021 )
Rsquare= 0.046   (max possible= 0.999 )
Likelihood ratio test= 10.63  on 1 df,   p=0.001
Wald test            = 10.09  on 1 df,   p=0.001
Score (logrank) test = 10.33  on 1 df,   p=0.001
```

Notice that the sex variable is encoded as a numeric vector(male:1, female:2). The Cox model summary gives the hazard ratio (HR) for the second group relative to the first group, in this case female vs male. The regression coefficients coef = -0.5310 means that female has a lower hazard (risk of death), and thus a better prognosis. The exponential coefficients e^{β} = 0.59, also known as hazard ratios, gives the effect size of covariates. We can tell being female reduces the hazard by a factor of 0.59, or 41%.

3 Resampling Procedures: Bootstrap

3.1 Confience Intervals for the Survival Function

In this section, we'll compare different versions of 90% confidence intervals for the survival function S(t) at the median time (50th percentile).
The five different methods of constructing CIs are:

1. CI from R Rusult
2. (bootstrap) CI by using Formula $\hat{\beta} \pm 1.645 se(\hat{\beta})$
3. (bootstrap) Percentile-Bootstrap CI
4. (bootstrap) Bias-Corrected Bootstrap CI
5. (bootstrap) Highest Probability Density CI

3.1.1 CI from R Rusult

From the result of the survival function S(t), the median is t = 310 days, and the survival at time t is 0.495. Then, we can also get the 90% confidence interval for t = 310.

```
Call: survfit(formula = surv ~ 1, data = cancer, conf.int = 0.9)


 time n.risk n.event survival std.err lower 90% CI upper 90% CI
  310     85     107    0.495  0.0352         0.44        0.557
```

The 90% confidence interval for the median time t = 310 is

$$[0.44, 0.557]$$

3.1.2 CI by using Formula

In package *Hmisc*, it provides a function ***bootkm(S, q, time, B)***, which bootstraps Kaplan-Meier estimate of the probability of survival to at least a fixed time (times variable) or the estimate of the q quantile of the survival distribution.
S: a Surv object for possibly right-censored survival time.
q: quantile of survival time, default is 0.5 for median.
time: time at which to compute survival estimates.
B: number of bootstrap repetitions.

Thus, by the following code, we can get a list of 10000 survival estimates at median time 310.

$$q = bootkm(surv, time = 310, B = 10000)$$

Then, we can calculate the mean and standard error of 10000 bootstrap observations, and use the formula to caculate the confident interval.

$$[LCL, UCL] = [\hat{q} - 1.645se(\hat{q}), \hat{q} + 1.645se(\hat{q})]$$

and we can get the 90% bootstrap CI for the median time t = 310 by using formula is

$$[0.4363, 0.5529]$$

3.1.3 Percentile-Bootstrap CI

For the Percentile-Bootstrap-CI method, we use 10,000*0.05 = 500th percentiles as the lower confidence interval, and use 10,000*0.95 = 9500th percentiles as the upper confidence interval.

```
       5%       95%
0.4349957 0.5521079
```

In this method, we can get the 90% Percentile-Bootstrap CI for the median time t = 310 is

$$[0.4350, 0.5521]$$

3.1.4 Bias-Corrected Bootstrap CI

The Acceleration and Bias-Correction (or BCa) method improves on the percentile method by adjusting the percentiles $(\tilde{\theta}(1-\alpha/2), \tilde{\theta}(\alpha/2))$ chosen from the bootstrap sample above.
The bias factor is

$$z_0 = \phi^{-1}(p_0)$$

where

$$p_0 = B^{-1} \sum I(\tilde{\theta}_i < \theta_n)$$

indicates the proportion of the bootstrap estimaters $\tilde{\theta}$ less than θ_n.
The acceleration factor measures the rate of change in σ_{θ_n} as a function of θ.

$$a_0 = \frac{\sum_{i=1}^{B} (\tilde{\theta}^* - \tilde{\theta}_i)^3}{6\left(\sum_{i=1}^{B} \left(\tilde{\theta}^* - \tilde{\theta}_i\right)^2\right)^{3/2}}$$

where $\tilde{\theta}^*$ is the average of the bootstrap estimaters.
The $100(1-\alpha)\%$ BCa interval is

$$\left[\tilde{\theta}(q1),\, \tilde{\theta}(q2)\right]$$

where

$$q1 = \phi\left(z_0 + \frac{z_0 + z_{\alpha/2}}{1 - a_0(z_0 + z_{\alpha/2})}\right)$$

$$q2 = \phi\left(z_0 + \frac{z_0 + z_{1-\alpha/2}}{1 - a_0(z_0 + z_{1-\alpha/2})}\right)$$

From our result of the survival function S(t), the median is t = 310 days, and the survival at time t is 0.495. Thus $\theta_n = 0.495$.
And $\tilde{\theta}_i$ is the i_{th} bootstrap estimators. $\tilde{\theta}^*$ is the mean of the 10000 bootstrap estimators.

```
5.317461% 95.30417%
0.4365107 0.5528617
```

By which we can get the 90% Bias-Corrected Bootstrap CI for the median time t = 310 is the 5.3175% and 95.3041% percentiles, which is

$$[0.4365, 0.5529]$$

3.1.5 Highest Probability Density CI

Highest Posterior Density (HPD) confidence interval is the shortest confidence interval enclosing $(1-\alpha)\%$ of the posterior probability. The HPD is an interval in which most of the distribution lies, which means the minimum density of any point within that region is equal to or larger than the density of any point outside that region.
We can use the textbf*HPDinterval()* function in R to calculate it.

```
      lower     upper
q 0.4385311 0.5548309
attr(,"Probability")
[1] 0.9
```

In this method, we can get the 90% Percentile-HPD CI for the median time t = 310 is

$$[0.4385, 0.5548]$$

This method constructs the shortest confidence interval of liength 0.11439 comparing to the other four methods above.

3.2 Compare five CIs

Since the estimate of survival distribution at median survival time remains unknown, we used five methods to construct 90% CI. The general bootstrap method, instead of using the estimate for the original data, is based only on bootstrap resamples. It does not adjust for skewness in the bootstrap distribution. When bootstrap resamples do not commit normal distribution, we can manually find lower and upper 5% resamples to construct 90% CI or use Bias-Corrected Bootstrap CI for a better CI approximation. Also, survival function program provides us with 90% CI based on original sampels and HPD-Interval method aims at finding the shortest 90% CI. Below is a summary table for five versions of CIs.

Method	**Confidence Interval**
R Result	[0.440, 0.557]
Bootstrap	[0.436, 0.553]
Percentile-Bootstrap	[0.435, 0.552]
Bias Correction(BC)	[0.437, 0.553]
HPD	[0.439, 0.555]

Table 2: CI Result

According to the summary table, five CIs are quite similiar to each other. Also, HPD-Interval does give us shortest 90% CI among five.

Furthermore, in order to have a better understanding of these results, we draw resamples distribution curve along with five CIs.

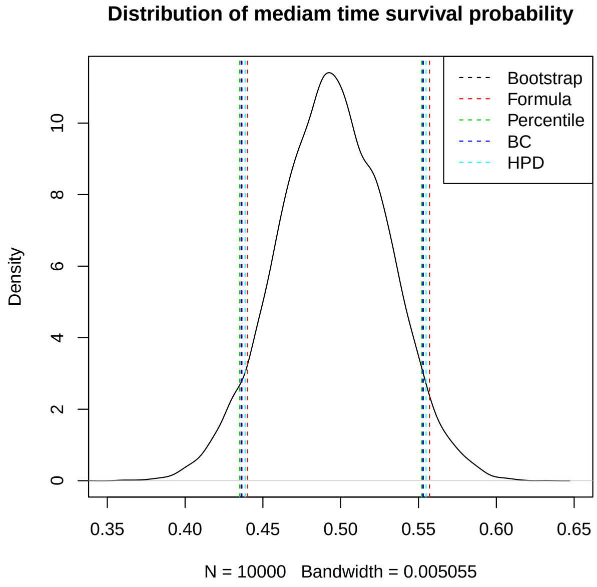

Figure 3: Resamples Distribution and Confidence Intervals

It can tell that resamples distribution has skewness in the middle, but with normal-like tails. In order to test its normality especially in tails distribution, we perform Anderson-Darling normality test, which is more sensitive to tails distribution than other normality tests.

```
        Anderson-Darling normality test


data:  q
A = 0.49964, p-value = 0.2091
```

According to A-D test result, p-value is 0.2091 so that the hull hypothesis that resamples distribution is normal can not be rejected. As we know, confidence interval is sensitive to tails distribution. Since bootstrap resamples commit normal distribution, it well explains the similarity of five CIs.