Lab 2 - Kaplan-Meier and Nelson-Aalen estimators - Simulations.

Introduction

This is a simulation lab. That is, I am expecting you to play around with some basic concepts in a controlled setting. You can choose the parameters and setting (shape of the distribution of the survival times, distribution and treatment of censoring, sample size). We will start with the larynx data and move on to simulated data.

1 Revisiting the Larynx data set.

This week we talked about confidence intervals for survival and cumulative hazard. Let's start the lab by adding confidence intervals to the estimated survival for the whole data set. Remember that these intervals are *pointwise.* Still, I expect you to use the intervals to make inferences about survival. In this lab we'll focus on the survival of all cancer patients, and in the upcoming lab we'll distinguish between patients with different diagnosis. However, if you are curious you can apply the commands and methods below to subsets (strata) of the data just to get a preview of what's to come.

We will begin with the Kaplan-Meier estimates;

```
surv.obj<-Surv(larynx$time,larynx$status)
```

First create a survival data object. Please check what `surv.obj` contains. Comment.

```
sf<-survfit(Surv(larynx$time,larynx$status)~1,type='kaplan-meier')
```

Look at the summary of the survival fit.

```
print(summary(sf))
```

The estimated error (standard error of the survival function) is based on Greenwood's formula (see book and lecture notes).

Now we are ready to set up confidence intervals for the survival. The default confidence interval is on the log-scale, but let's start with the simple $\hat{S(t)} \pm z_{1-\alpha/2} SE(\hat{S(t)})$.

```
sf<-survfit(Surv(larynx$time,larynx$status)~1,type='kaplan-meier',conf.type='plain')
plot(sf)
```

Please comment on the shape of the survival function, and the uncertainty associated with the estimate. What is you apply the above commands to a smaller subset of the data (`larynx.sub<-larynx[sample(seq(1,dim(larynx)[1]),frac),]` Pick the number of observations `frac` for the data subset yourselves.)

In class we talked about other estimators of both survival and the standard error. Let us compare the Nelson-Aalen estimator to the Kaplan-Meier estimate above.

```
sf2<-survfit(coxph(Surv(larynx$time,larynx$status)~1),type='aalen',conf.type='plain')
plot(sf2)
```

For this, rather large, data set, there isn't much of a difference between the two estimates. (Check what happens on a smaller subset of the data.)

```
plot(sf$time,sf$surv,type='s',ylim=c(0,1))
```

```
lines(sf2$time,sf2$surv,type=’s’,col=2)
```

Sometimes it’s preferable to look at the survival on a log-scale (easier to ’test’ if an exponential distribution would suffice to describe the data well).

```
sf2<-survfit(coxph(Surv(larynx$time,larynx$status)∼1),type=’aalen’,conf.type=’log’)
plot(sf2)
```

You may also consider looking at the estimated cumulative hazard rather than the survival.

```
plot(survfit(coxph(Surv(larynx$time,larynx$status) 1)),fun=’cumhaz’)
```

Please comment on all that you have done so far. What can you tell me about the larynx cancer data? Please use your choice of graph to deduce what the median survival time is (and give a rough idea of what the confidence interval is for median survival).

2 Censoring - impact on estimation

We talked briefly in class about what we would expect to happen if we ignored censoring when estimating survival. In this section I ask you to simulate a few scenarios and study the estimates you obtain ignoring censoring.

Let’s first simulate a survival data set. We can use a log-normal distribution to keep things simple. You should pick the sample size (n), and the mean and standard deviation of the event distribution and the censoring times (u,v,u2,v2). Try a couple of different values.

```
Y<-rnorm(n,mean=u,sd=v)
simtime<-exp(Y)
simcens<-exp(rnorm(n,mean=u2,sd=v2))
whoisbigger<-sign(simcens-simtime)
```

The last command checks to see if the censoring time (simcens) occurred before, or after, the real event time. If simtime is less than simcens, we observe the real event time, if simtime exceeds simcens, we observe the censoring time.

Let’s now create a censored data set.

```
simstatus<-(whoisbigger+1)/2
```

Confirm that this creates an indicator vector that is equal to 1 for the real event times.

```
obstime<-simtime*simstatus+simcens*(1-simstatus)
```

Using the survival functions from the previous section, ’analyze’ the observed survival data you created (`obstime`). Change the mean and standard deviation of the event distribution and/or the censoring distribution. Discuss your findings.

What if you ignore censoring (that is, do not use the information in `simstatus` appropriately)? As an example, set all the status indicators to 1 when you estimate the survival. Comment on the results.

What if you remove all censored data prior to estimation? Comment?

In all of the above, please also use the estimated survival curves to estimate the median survival time. Comment on the different estimates obtained. What is the impact of treating censored observations as real event times? What is the impact of dropping the censored observations from the data?

3 Optional - correlated censoring and survival times.

For the PhD students in the class, this is not optional.

Create a correlation structure between the survival and censoring times.
`Sigma<-matrix(sig*c(1,rho,rho,1),2,2)`
Here, you pick `sig` (the overall, identical marginal covariance) and `rho` (the correlation). Try both positive and negative correlations `rho`, as well as small or large ones.
Generate survival and censoring times jointly;
`both<-exp(mvrnorm(n,c(u1,u2),Sigma))`
u1 and u2 are the two means for survival and censoring, and n is the sample size. Now you can proceed as above; the first column of `both` contains the survival times (`both[,1]`), and the second column the censoring times.

Please comment on your findings. What is the impact on estimating survival/median survival time if the censoring times and survival times are positively/negatively correlated? Explain why you get these results.

Summary

Please write up a short report outlining your work in this lab. Include only the relevant figures.