Introduction to Biostatistics (171:161)
Breheny

Lab #15

In lecture this week we've looked at survival analysis. Today's lab will focus on Kaplan-Meier survival curves and the log-rank test. We will examine these using SAS and `R` on the aplastic anemia dat set. The data set contains five variables:

- `Trt`: Whether the patient received Methotrexate (MTX) or Methotrexate and cyclosporine (MTX+CSP).
- `Time_GVHD`: Time until graft-versus-host disease. Measured in years.
- `Status_GVHD`: What happened at the end of `Time_GVHD`. The patient was either censored (0) or got graft-versus-host disease (1).
- `Time`: Time until death. Measured in years.
- `Status`: What happened at the end of `Time`. The patient was either censored (0) or died (1).

In both SAS and `R`, before you can carry out an analysis, you need to specify how to combine the time on study and censoring information into the time-to-event outcome you wish to analyze. Somewhat confusingly, SAS has you specify which events are censored, while `R` has you specify which events are observed (*i.e.*, which events are *not* censored).

The procedure in SAS that performs basic survival analysis procedures is called `PROC LIFETEST`. To run `PROC LIFETEST`, you need a `TIME` statement that tells SAS which variable records the time on study and how the censoring events are recorded in the data set. The package in `R` that performs survival analysis is called `survival`; to use its functions, you need to set up a survival object using the function `Surv`.

SAS:

```
PROC LIFETEST DATA=anemia;
  TIME Time*Status(0);
RUN;
```

`R`:

```
require(survival)
S <- Surv(Time,Status!=0)
fit <- survfit(S~1)
plot(fit, ylab = "Probability", xlab = "Time")
```

The above code calculates the Kaplan-Meier survival function for the study as a whole. Note that in SAS, the syntax of the time statement is `TIME` followed by the name of the time on study variable, then an asterisk, then the name of the variable that contains the censoring information, and finally the values of the variable that correspond to censoring. Here, 0 corresponds to censoring. In `R`, the syntax is similar for the `Surv` function, although we need to specify which events are *not* censored (in `R`, `!=` means "not equal). Equivalently, we could have created the survival object with `S <- Surv(Time,Status==1)`.

Looking at the output, take note of the following:

- SAS refers to the Kaplan-Meier estimates as the “Product-Limit” estimates. This is another common name for the estimation procedure we talked about in class.
- By default, SAS prints out a big list of all the subjects, sorted by time on study, the estimated survival and failure functions (the failure function is just the 1 minus the survival function), and the number at risk (“Number left”). In you would like this information from `R`, type `summary(fit)`.
- At the end of the list, SAS prints some summary statistics, such as the median survival time, the total number of subjects, the total number who failed, and the total number who were censored.
- In `R`, the formula `S~1` means that everyone is in the same group – as opposed to a formula like `S~Group`, which we have seen for functions like `t.test`.

However, the overall survival of the patients isn’t what we want to know. We want to know how the survival differs between the two treatment groups (if at all). To obtain separate survival estimates, we need to include a `STRATA` statement in SAS, and use the `survdiff` function in `R`:

SAS:

```
PROC LIFETEST DATA=anemia;
  TIME Time*Status(0);
  STRATA Trt;
RUN;
```

`R`:

```
survdiff(S~Trt)
fit <- survfit(S~Trt)
plot(fit, ylab = "Probability", xlab = "Time")
```

The above code now provides separate estimates for the `MTX` and `MTX+CSP` groups. In addition, we perform a log-rank test for differences in the survival function between the two groups (in SAS, it’s at the end, under “Test of Equality over Strata”). Note that the *p*-value here is .16.

NOTE: The plot function in the `survival` package in `R` is not the greatest. In particular, it doesn’t label anything for you. Fully explaining all the plotting options in `R` is a bit beyond the scope of this course, but here is an example of their use:

```
plot(fit,mark.time=FALSE,col=c("gray","slateblue"),lwd=3,
     xlab="Time on study (Days)",ylab="Probability of survival")
text(900,.4,"MTX")
text(900,.9,"MTX+CSP")
```

Note that by default, SAS and `R` add little tick marks onto the Kaplan-Meier plots at all of the times for which censoring occurred. This occasionally provides useful information, but can also clutter the plot. To remove them, one can specify `mark.time=FALSE` in `R` (see above) or `NOCENSOR` in SAS:

```
PROC LIFETEST DATA=anemia PLOTS=S(NOCENSOR);
  TIME Time*Status(0);
  STRATA Trt;
RUN;
```

Finally, in SAS one can obtain confidence intervals and place them on the Kaplan-Meier curve by adding a `CL` option (this works with or without the `NOCENSOR` option):

```
PROC LIFETEST DATA=anemia PLOTS=S(NOCENSOR CL);
  TIME Time*Status(0);
  STRATA Trt;
RUN;
```

SAS does a rather beautiful job here of coloring the two confidence bands to illustrate where they overlap and where they do not. Note that, as discussed in class, the intervals get wider as more and more people are censored.

NOTE: Once again, it is unfortunate that the `survival` package does not provide a better plotting function (if you want to see something really ugly, type `plot(fit,conf=TRUE)`). A much better plotting function called `survplot` is available in the `rms` package. If you are interested, type `install.packages("rms")`, then `require(rms)`, then `survplot(fit)`. Much better!

Of course, there were two outcomes that the study investigated. In addition to survival, the study looked at graft-versus-host disease (GVHD). To perform the same analysis, but for GVHD, we submit:

SAS:

```
PROC LIFETEST DATA=anemia;
  TIME Time*Status(0);
  STRATA Trt;
RUN;
```

R:

```
S <- Surv(Time_gvhd,Status_gvhd)
survdiff(S~Trt)
fit <- survfit(S~Trt)
plot(fit) ## or survplot(fit)
```

Note that the p-value here is .01. The study finds little evidence for a difference in overall survival, but fairly strong evidence for a difference in graft-versus-host disease.

As for the plot, notice that the interesting part of the curve is cramped into the far left-hand side of the plot. Unfortunately, SAS gives you no way to change the range of horizontal axis (hopefully a future version of SAS will remedy this problem, but I've been saying this for a few versions now). In R, you can type `plot(fit,xlim=c(0,100))` to only plot the portion of the curve between 0 and 100 days; you can then add labels as above, or use `survplot`, which works with `xlim` as well.