Survival Analysis - a *brief* overview

Survival analysis is concerned with exactly that. For example, we may be interested in the time to death after a the onset of some disease, or the time to death following a treatment for cancer (the treatment being one that extends life). Although survival analysis doesn't have to be restricted to biological events (e.g., we might be interested in the time until an engine breaks down), it is most often used in medicine and biology.

We'll discuss two components of survival analysis: 1) Survival curves (e.g. Kaplan-Meier estimates) and the Cox proportional hazards model. For neither of these will be go into any details.

1 Basic survival analysis (including curves)

It's probably easiest to do an example. We'll use an example based on a web page (`http://sthda.com/english/wiki/survival-analysis-basics#kaplan-meier-survival-estimate`) that illustrates the basics (not all of this is useful for us).

Before we begin, we need to install two R packages: `survival` and `survminer`. The first of these provides the basic survival analysis framework (think "math"), while the second lets us visualize a lot of what the first package does. To install these in R from the command line you can do:

```
install.packages(c("survival", "survminer"))
```

If you prefer, you can click on the `Packages` tab in RStudio (lower right window), click on `Install` and type the names of the packages separated by a comma in the pop-up box (make sure `Install dependencies` is checked).

If you're using Windows or MacOS you shouldn't have any issues. If you're using Linux, this may result in all kinds of missing libraries and other issues. In this case you're stuck messing around with finding the missing libraries/packages and installing them on either Linux or R as needed (annoying, a bit time consuming, but it does work).

Once you have the packages installed, you need to load them (R doesn't do this automatically):

```
library("survival")
library("survminer")
```

There are a large number of data sets in the `survival` package, but to keep things simple

we'll stick with the one in the example mentioned above. Let's look at it:

```
head(lung)
```

The `head` command tells R to only print the first 6 lines of the dataset - that's so you can *look* at the data without having to scroll through pages of output. You should get back the following:

```
  inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1    3  306      2  74   1       1        90       100      1175      NA
2    3  455      2  68   1       0        90        90      1225      15
3    3 1010      1  56   1       0        90        90        NA      15
4    5  210      2  57   1       1        90        60      1150      11
5    1  883      2  60   1       0       100        90        NA       0
6   12 1022      1  74   1       1        50        80       513       0
```

Let's look at these variables (you can also do `help(lung)` and then click on the link that pops up):

inst	A code for the institution at which the patient was observed.
time	The time in days the patient survived or was censored.
status	The censoring status (1 = censored, 2 = dead).
age	The age of the patient in years.
sex	1 = Male, 2 = Female.
ph.ecog	ECOG performance (0 - 4; 0 = asymptomatic, 4 = bedbound).
ph.karno	A performance score (Karnofsky) rated by physician (0 = bad, 100 = good).
pat.karno	Same as above, but rated by patient.
meal.cal	The calories consumed at meals.
wt.loss	The amount of weight (lbs.) lost in the last six months.

That might seem like a lot of variables, but some of the other data sets (e.g., `colon`) have a lot more.

The first thing you may notice is that there are several references to *censoring* above. Many data sets used for survival analysis include variables that are censored. This means that, for example, the patient survived until the end of the study. If the patient survived, we *don't known* how much longer the patient may have survived - they may still be alive! But that means we don't know if or when the patient died. All we know is that the patient survived at least as long as the study.

In other words, if the `status` variable indicates a `1` above, it means the patient survived through the end of the study - we don't have information about *time to death*. This is what we mean by censoring. Survival analysis will take account of censored data - in particular,

the *Kaplan-Meir* method of calculating survival curves (more below) can be used if you have censored data (it can also be used with non-censored data).

Let's continue our example by calculating the emphmedian survival time by sex. We should be familiar with some of the syntax below.

```
medsurvival <- survfit(Surv(time, status)   sex, data = lung)
```

The `survfit` function helps to generate the survival curves. The `Surv` option combines the time in days with the censoring status into a single variable which the `survfit` function can interpret.

If we look at the results of the above (use `print(medsurvival)` we get the following results:

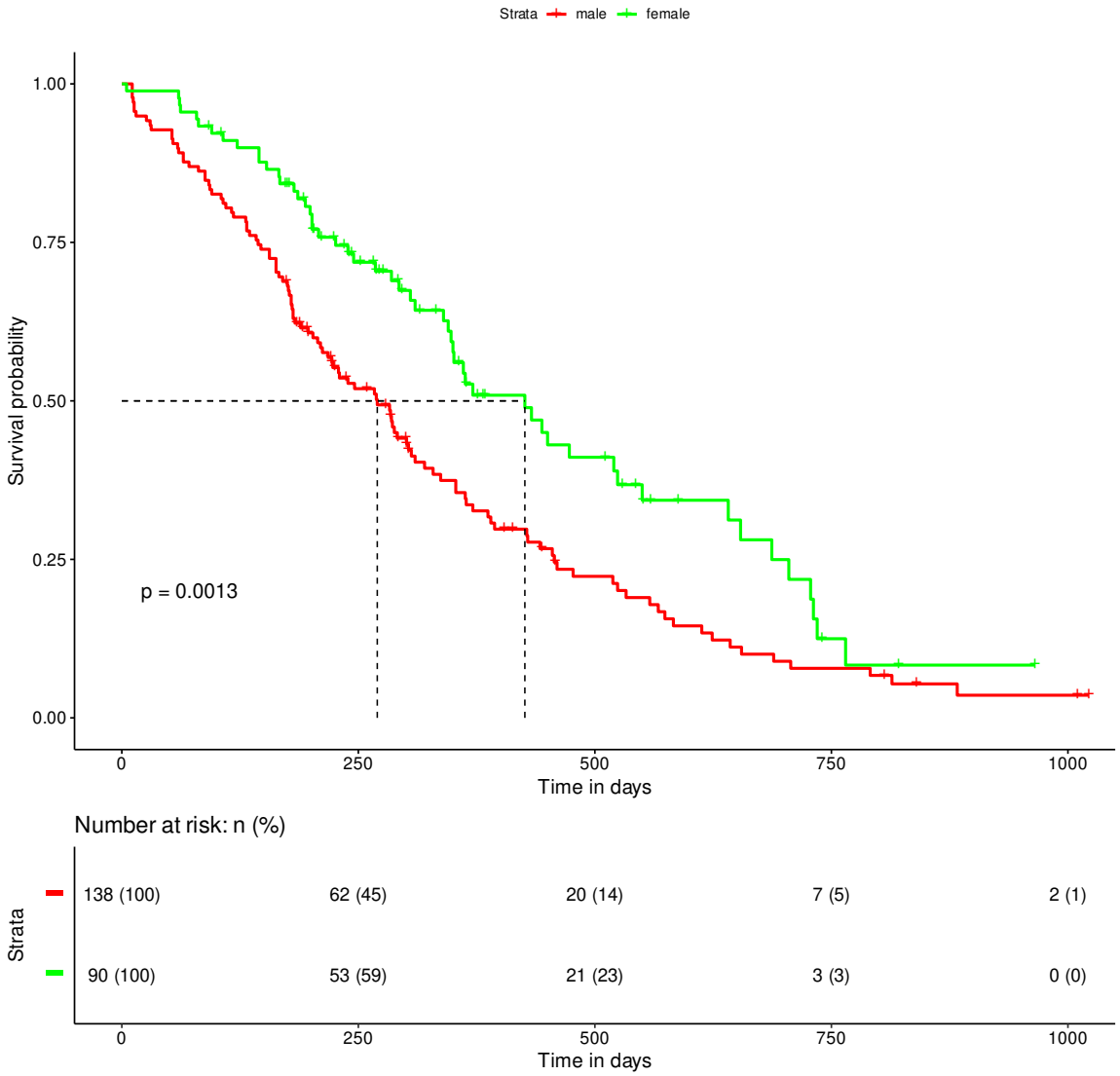
```
Call: survfit(formula = Surv(time, status) ~ sex, data = lung)

        n events median 0.95LCL 0.95UCL
sex=1 138    112    270     212     310
sex=2  90     53    426     348     550
```

This gives us the sample sizes, number of deaths (`events`), the median survival time, and confidence limits for the median survival time. If you want, you can get a little more information (e.g. the means) using `summary(cfit)$table`.

But let's now see what it all looks like. To do this, we'll plot the survival curves as follows:

```
ggsurvplot(
  cfit,                       # survfit object with calculated statistics.
  pval = TRUE,                # show p-value of log-rank test.
  conf.int = FALSE,           # show confidence intervals.
  # conf.int.style = "step",  # customize style of confidence intervals
  xlab = "Time in days",      # customize X axis label.
  # ggtheme = theme_light(),  # customize plot and risk table with a theme.
  risk.table = "abs_pct",     # absolute number and percentage at risk.
  risk.table.y.text.col = T,  # color risk table text annotations.
  risk.table.y.text = FALSE,  # show bars instead of names in text annotations
                              # in legend of risk table.
  # ncensor.plot = TRUE,      # plot the number of censored subjects at time t
  surv.median.line = "hv",    # add the median survival lines.
  legend.labs =
    c("male", "female"),      # change legend labels to something reasonable.
  palette =
    c("red", "green")         # pick your colors.
)
```

There are actually many more options (!), but the ones above are the ones most important for us. Most of them should be explained by the comments above. The ones commented out are ones you might find interesting if you try them. In any case, this gives us the following graph:



The curves show the the survival probability based on days. At 0 days, everyone is still alive, but at 1,000 days very few people are still alive (it is lung cancer). You can also see the median lines and you can notice that the median survival time for men is less than for women (the printout above also showed us this). The p-value is a test for the difference in survival of the two curves. It's called the log-rank test, and is basically calculated using the χ^2 goodness of fit procedure. From this we see that the survival for men and women is significantly different (assuming any reasonable level of α).

Finally, at the bottom we see the number of people still alive at the various time intervals (with % given in parenthesis).

The exact calculations for these curves are based on simply the number alive at a given time. The Kaplan-Meir method then incorporates the information on censored individuals. If you're interested in the calculations (they're not complicated), you can refer to on-line searches. Or a good reference text is *Survival Analysis - A Self-Learning Text* by *Kleinbaum and Klein*. The latest edition as of 2023 is the third.

2 The Cox proportional hazards model

We don't have the time to delve into the details here, but the Cox proportional hazards model basically turns what we learned above into a (multiple) regression. By doing this, we can analyze which variables might be important in predicting survival. We'll stick with the example above. Remember that there are a bunch of other variables in our data set (ECOG performance, weight loss, etc.). Let's see how we can incorporate this information into our survival analysis (we won't discuss the math behind all of this).

We'll start with a simple analysis (similar to above) looking at just sex:

```
lungcox <- coxph(Surv(time, status)   sex, data = lung)
lungcox
```

Which gives us:

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

       coef exp(coef) se(coef)      z        p
sex -0.5310    0.5880   0.1672 -3.176 0.00149

Likelihood ratio test=10.63  on 1 df, p=0.001111
n= 228, number of events= 165
```

This is very similar to what the log-rank test gave us above when we looked at our survival curves. Note the negative (−) coefficient for `sex`. This tells us the risk (=*hazard*) goes *down* as we go from males to females (women did better in this study). In other words, a (−) coefficient tells us the *hazard* goes down. A (+) coefficient tells us the hazard increases.

As it turns out, the coefficients are actually *hazard ratios*. They compare the hazards of the various categories (in this case the two levels of sex). This is why it's called the Cox proportional *hazards* model. It is also an important assumption that this ratio is constant - the ratio doesn't change with time. We don't have the time to deal with checking this

assumption, but if you start using the Cox proportional hazards model you should look into this!

So where are we? We haven't really done anything new yet. Let's try using the Cox proportional hazards model and including some of other variables described above:

```
lungcox2 <- coxph(Surv(time, status)   age + sex + ph.ecog + ph.karno + wt.loss,
data = lung)
summary(lungcox2)
```

This time we get the following result:

```
Call:
coxph(formula = Surv(time, status) ~ age + sex + ph.ecog + ph.karno +
    wt.loss, data = lung)

  n= 213, number of events= 151
   (15 observations deleted due to missingness)

              coef exp(coef)  se(coef)      z Pr(>|z|)
age       0.015157  1.015273  0.009763  1.553 0.120538
sex      -0.631422  0.531835  0.177134 -3.565 0.000364 ***
ph.ecog   0.740204  2.096364  0.191332  3.869 0.000109 ***
ph.karno  0.015251  1.015368  0.009797  1.557 0.119553
wt.loss  -0.009298  0.990745  0.006699 -1.388 0.165168
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          exp(coef) exp(-coef) lower .95 upper .95
age          1.0153     0.9850    0.9960    1.0349
sex          0.5318     1.8803    0.3758    0.7526
ph.ecog      2.0964     0.4770    1.4408    3.0502
ph.karno     1.0154     0.9849    0.9961    1.0351
wt.loss      0.9907     1.0093    0.9778    1.0038

Concordance= 0.64  (se = 0.026 )
Likelihood ratio test= 33.53  on 5 df,   p=3e-06
Wald test            = 32.27  on 5 df,   p=5e-06
Score (logrank) test = 32.83  on 5 df,   p=4e-06
```

We should be familiar with how to interpret some of this by now. The coefficient for sex is still negative, and sex is still significant. The only other variable that is significant is `ph.ecog`, or ECOG performance. If you remember it is coded from 0 to 4, with 4 being the worst. It shouldn't be a surprise that the coefficient is nice and positive - telling us the higher the ECOG rating, the worse for the patient.

Let's not worry about the second part of the printout (the exponentiated coefficients). But

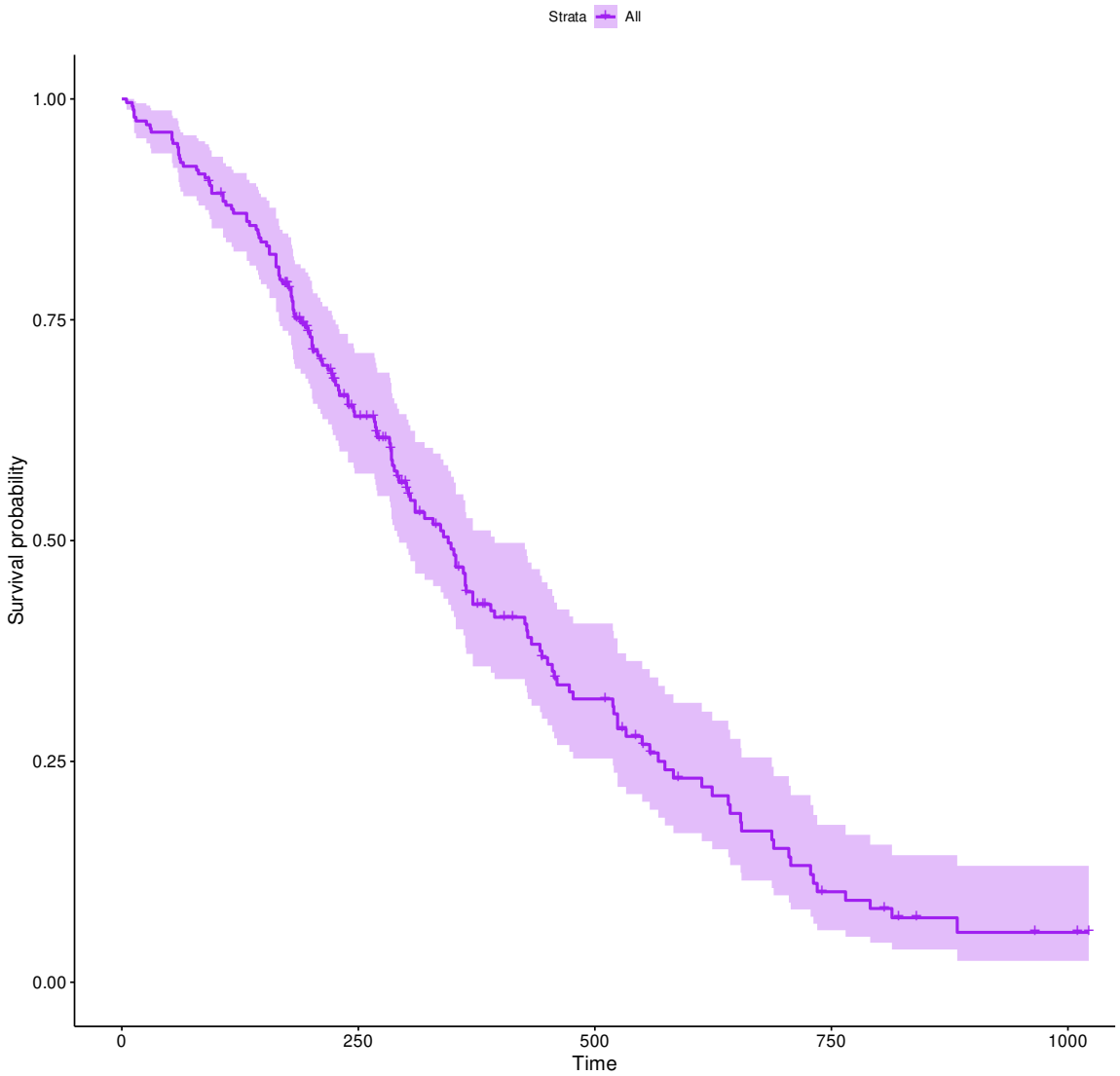
if we go the very bottom of the printout we see three tests (`Likelihood ratio, Wald, and Score (logrank)`). These tests evaluate the *overall* significance of the model. In other words, is the model with all the variables we used significant? R prints three different test results because some people prefer one test or another (but it's good if they all agree!!).

You can think of this a bit like the overall ANOVA F test, followed by Tukey's for the pairwise differences. These test tell us the model is significant, which means that sex and ECOG are significant (if none of these tests had been significant we could *not* say sex and ECOG are significant).

It's important to notice that the coefficients (and significance) can change depending on which variables we put in the model. If, for example, we substitute the patient rated Karnofsky performance scale for the physician rated performance scale the coefficients for sex and ECOG change somewhat, and we pick up significance for the patient rated performance scale. (Try it - substitute `pat.karno` for `ph.karno` in the command above). You need to be just a little careful in interpreting the coefficients because the model depends on everything we put into it!

Finally, we can also plot this model, similarly to what we did above. We'll only get one curve this time that shows the overall survivorship with all the variables that we used in the model:

```
ggsurvplot(survfit(lungcox2), palette = "purple", data = lung)
```



This time the confidence intervals are plotted (we didn't plot them before because the obscure the individual curves just a little).

There is obviously a lot more to survival analysis. If you're interested, the text mentioned above is a very readable and should provide a lot more information.