MA256 Lesson 28 Survival Analysis and Censored Data

(Much of the material below is from ISLR2 https://www.statlearning.com/, often taken verbatim.)

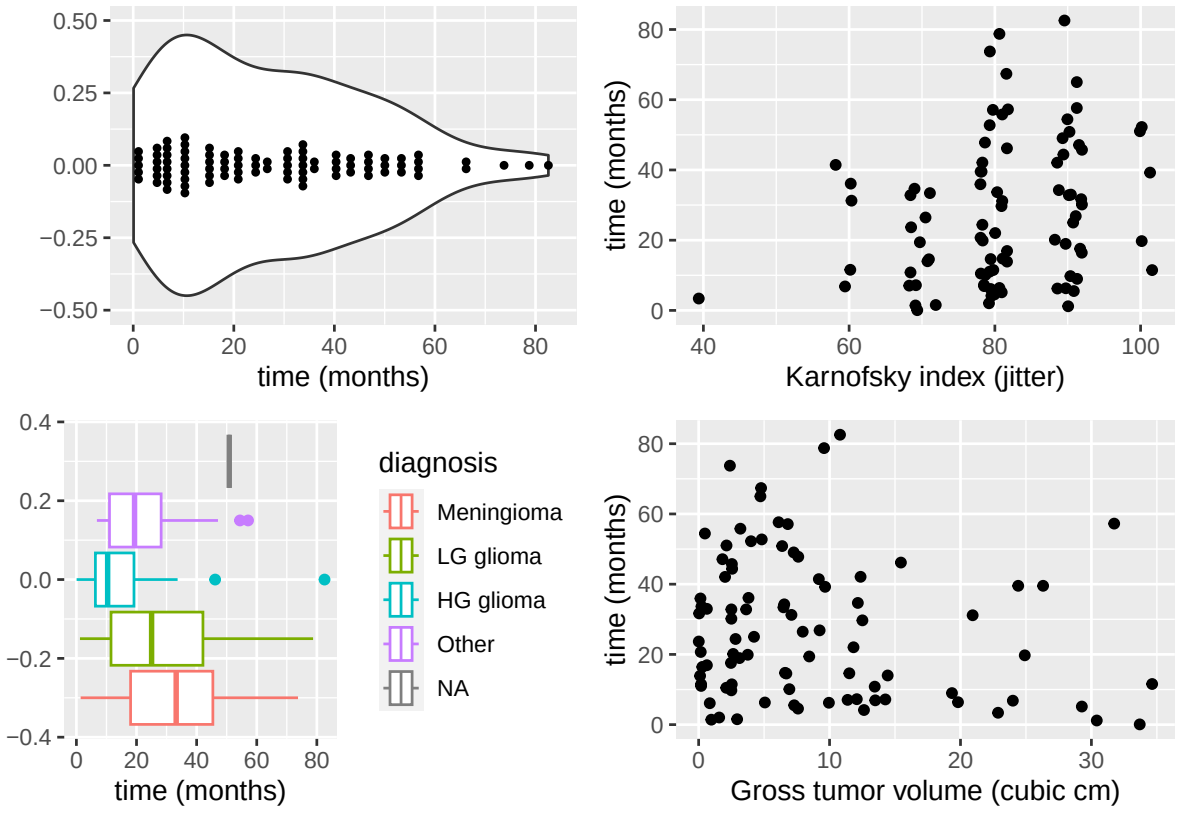
Brain Cancer Example

Suppose we have conducted a five-year medical study, in which the patients have been treated for cancer. We would like to fit a model to predict patient survival time, taking into account various features.

First, we perform some EDA. Here we show:

1) A violin plot with the distribution of survival times (months) for those with cancer;
2) Survival time vs. the Karnofsky index (a measurement of the functional impairment for a patient. See: http://www.npcrc.org/files/news/karnofsky_performance_scale.pdf);
3) The survival times for various types of diagnoses ("Meningioma", "LG glioma", "HG glioma", and "Other");
4) The survival time for various gross-tumor sizes (in cm^3).

```
## Bin width defaults to 1/30 of the range of the data. Pick better value with
## 'binwidth'.
```



1) What do we see in these figures?

2) We can even conduct a linear regression on these variables:

```
# bc.lm <- bc %>% lm(time ~ XXXX + XXXX + XXXX, data = .)
# ### summary(bc.lm)
# anova(bc.lm)
```

3) Which variable did we ignore here? (hint: type `? BrainCancer` and read about the variables in the data set.)

4) By ignoring this variable identified in 2), what are some potential pitfalls that we may encounter?

Censored data

To help with this issue, we introduce the idea of censored data. We suppose that we know the true *survival time*, T, as well as a true *censoring time*, C. The survival time represents the time at which the event of interest occurs: for instance, the time at which the patient dies, or the customer cancels his or her subscription. By contrast, the censoring time is the time at which censoring occurs: for example, the time at which the patient drops out of the study or the study ends.

We then define the random variable:

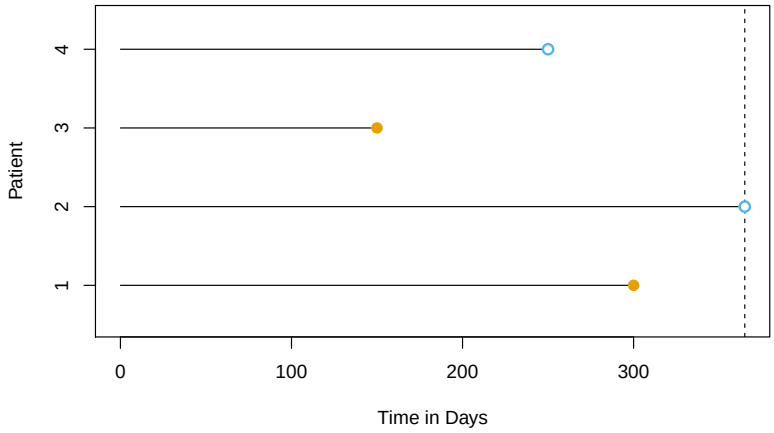
$$Y = \min(T, C) \tag{1}$$

and the status indicator:

$$\delta = \begin{cases} 1 & \text{if } T \leq C \\ 0 & \text{if } T > C. \end{cases} \tag{2}$$

where $\delta = 1$ indicates that we observe the true survival time and $\delta = 0$ indicates that we observe the censoring time. We record the n observations as the pair (Y, δ), which we can denote as $(y_1, \delta_1), \ldots, (y_n, \delta_n)$.

5) Consider the following picture, representing $n = 4$ patients for a 365-day follow up study after the cancer study. Write out the 4 observations using notation described above and describe what you are seeing:



6) Update the figures above to take into account the fact that some of the observations are censored. What do we see now?

Kaplan-Meier Survival estimator

The survival curve or survival function is defined as

$$S(t) = P(T > t) \tag{3}$$

7) For this statement: a) Write out what this means using words; b) write out an integral for this statement; and c) relate $S(t)$ to the CDF, $F(t)$.

8) Now we will look at a few ways to estimate Eq. (3).

a) Estimate $S(20) = P(T > 20)$ using the `bc$time` vector by counting the number of people who had survived past this time. What is one issue with this?

```
# sum(bc$time > XXXX) / XXXX
```

b) We can also estimate $S(20) = P(T > 20)$ by calculating the proportion of patients where $Y > 20$ for those who were not censored by time $t = 20$.

```
# # 71 not censored by time t=20
# not.cens.by.t.20 <- XXXX - sum(bc$status[bc$time <= XXXX] == 0)
# # 48 survived past Y > 20
# at.risk <- length(bc$time[bc$time > 20])
#
# at.risk / not.cens.by.t.20
```

some details...

A possible solution for these two issues is to calculate the Kaplan-Meier survival curve. To do so we define the following: - $d_1 < d_2 < \ldots < d_K$: denotes the K *unique* death times among the non-censored patients.

- q_k: denotes the number of patients who *died at time* d_k.
- r_k: denotes the number of patients still alive and in the study just before d_k.

By the total law of probability

$$\begin{aligned} P(T > d_k) =& P(T > d_k | T > d_{k-1}) P(T > d_{k-1}) \\ &+ P(T > d_k | T \leq d_{k-1}) P(T \leq d_{k-1}) \end{aligned}$$

but since $d_{k-1} < d_k$ we have $P(T > d_k | T \leq d_{k-1}) = 0$ so we have:

$$\begin{aligned} S(d_k) &= P(T > d_k) \\ &= P(T > d_k | T > d_{k-1}) P(T > d_{k-1}) \\ &= P(T > d_k | T > d_{k-1}) S(d_{k_1}) \\ &= P(T > d_k | T > d_{k-1}) \times \ldots \times P(T > d_2 | T > d_2) P(T > d_1) \end{aligned}$$

We can estimate the probability that a person will survive after time j, given they have survived until time $j-1$ with

$$\hat{P}(T > d_j | T > d_{j-1}) = (r_j - q_j)/r_j$$

We now have the Kaplan-Meier estimator to the survival curve:

$$S(d_k) = \prod_{j=1}^{k} \left(\frac{r_j - q_j}{r_j} \right)$$

and for times t between d_j and d_{j+1} we set $\hat{S}(t) = \hat{S}(d_j)$, which gives the K-M curve a step-like shape.

9) Use the function `survfit()` to estimate and plot the K-M survival curve for the full dataset and then again stratifying by sex, accounting for the variable `status`. What is the estimate for S(20) using the K-M survival curve?

```
# attach(BrainCancer)
# fit.surv <- survfit(Surv(time, status) ~ XXXX)
# plot(fit.surv, xlab = "Months",
#     ylab = "Estimated Probability of Survival")
#
# #### get probability of survival for t = 20
# ind <- tail(which(XXXXXXXXX), n=1)
# fit.surv$surv[ind]
# #### detach(BrainCancer)
```

Log-rank test

10) The figure below shows the K-M survival curves for males and females. We can see that females tend to have a better survival rate up until about 50 months when they both level out. Interpret what you see.

```
# fit.sex <- survfit(Surv(time, status) ~ XXXX)
# plot(fit.sex , xlab = "Months", ylab = "Estimated Probability of Survival", col = c(2,4))
# legend("bottomleft", levels(sex), col = c(2,4), lty = 1)
```

11) How can we compare the two curves? Is there a formal test that we can use? Describe how we could use a two-sample t-test to compare the differences. What is a problem with this?

some details...

A solution to compare the two curves is to use the log-rank test statistic. To construct the log-rank test statistic, you create a 2 x 2 table for each unique date d_j,

	Group 1	Group 2	Total
Died	q_{1j}	q_{2j}	q_j
Survived	$r_{1j} - q_{1j}$	$r_{2j} - q_{2j}$	q_j
Total	r_{1j}	r_{2j}	r_j

where we break up the total number of patients who died at time j ($q_{1j} + q_{2j} = q_j$) and the number of patients who are at risk at time j ($r_{1j} + r_{2j} = r_j$), for groups 1 and 2, respectively.

Ultimately we arrive at the test statistic W (details in in the ISLR):

$$W = \frac{\sum_{j=1}^{K}(q_{1j} - E(q_{1j}))}{\sqrt{\sum_{j=1}^{K}(VAR(q_{1j}))}}$$

With a large enough sample size, the log-rank statistic has an approximate standard normal distribution. You can then calculate a p-value to the null hypothesis that is no difference between the two survival curves.

12) use the `survdiff()` function to calculate the log-rank statistic to compare the survival of males to females. What is your conclusion?

```
# logrank.test <- survdiff(XXXX ~ XXXX)
# logrank.test
```

Regression Models with a Survival Response (§11.5)

Now we will fit a regression model where the observations are of the form (Y, δ) (from (1) and (2)). We will also consider additional explanatory variables, $X \in \mathbb{R}^p$, which is a vector with p features.

Goal: Predict the true survival time, T. (but we only have Y, the minimum of T and C!!!)

We will use a similar idea to that presented in the K-M survival curve and use sequential construction.

But first... the hazard function... this is also known as the *hazard rate* (why?) or the *force of mortality* and is defined as:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t) \cap (T > t))/\Delta t}{P(T > t)} \\ &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t))/\Delta t}{P(T > t)} \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

With this we can move to the *proportional hazards assumption*:

$$h(t|x_i) = h_0(t) \exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right) \tag{4}$$

where $h_0(t) \geq 0$ is an unspecified function, known as the baseline hazard for a person with features (explanatory variables) $x_{i1} = \cdots = x_{ip} = 0$. The other term, $\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$ is called the relative risk for a person with feature vector $x_i = (x_{i1}, \ldots, x_{ip})^T$, relative to the feature vector $x_i = (0, \ldots, 0)^T$.

What does it mean that the baseline hazard function $h_0(t)$ is unspecified? Basically, we make no assumptions about its functional form. We allow the instantaneous probability of death at time t, given that one has survived at least until time t, to take any form. This means that the hazard function is very flexible and can model a wide range of relationships between the covariates and survival time. Our only assumption is that a one-unit increase in x_{ij} corresponds to an increase in $h(t|x_i)$ by a factor of $\exp(\beta_j)$.

Cox's proportional hazards model

We can use (4) and the "sequential in time'' idea that we used above. First we assume that an observation is uncensored ($\delta_i = 1$) and thus y_i is the actual failure time ($y_i = T_i$). The hazard function is defined as in (4) and the total hazard at time y_i for the at risk observations (those that have not failed yet) is given by:

$$\sum_{i':y_{i'} \geq y_i} h_0(y_i) \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right) \tag{5}$$

Thus we can calculate the probability of the ith observation will fail at time y_i is given by:

$$\frac{Eqn.(4)(y_i)}{Eqn.\ (5)} = \frac{h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}$$
$$= \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}$$

We can see that the baseline hazard function cancels out of numerator and denominator. We can use the partial likelihood (See ISLR) to estimate the β_j's.

13) Use the function `coxph()` to use the Cox proportional hazards model using `sex` as the only predictor. What is our conclusion?

```
# fit.cox <- coxph(Surv(time, status) ~ sex)
# summary(fit.cox)
```

14) How does our result from the Cox prop. haz. model compare to what we saw with the log-rank test? (check the `chisq` result from our log-rank model)

```
# logrank.test$chisq
```

15) Repeat the Cox prop. haz. model with additional explanatory variables: `sex, diagnosis, loc, ki, gtv, stereo`. Explain what you see. Does the negative coefficient make sense?

```
# fit.all <- coxph(XXXX ~ XXXXX)
# summary(fit.all)
```

16) Plot the various survival curves for each diagnosis category, while keeping the other explanatory variables fixed. Does this match the exploratory figures we plotted above?

```
# modaldata <- data.frame(diagnosis = levels(diagnosis),
#                         sex = rep("Female", 4),
#                         loc = rep("Supratentorial", 4),
#                         ki = rep(mean(ki), 4),
#                         gtv = rep(mean(gtv), 4),
#                         stereo = rep("SRT", 4)
# )
# survplots <- survfit(fit.all, newdata = modaldata)
#
# plot(survplots,
#      xlab = "Months",
#      ylab = "Survival Probability", col = 2:5)
# legend("bottomleft", levels(diagnosis), col = 2:5, lty = 1)
```