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

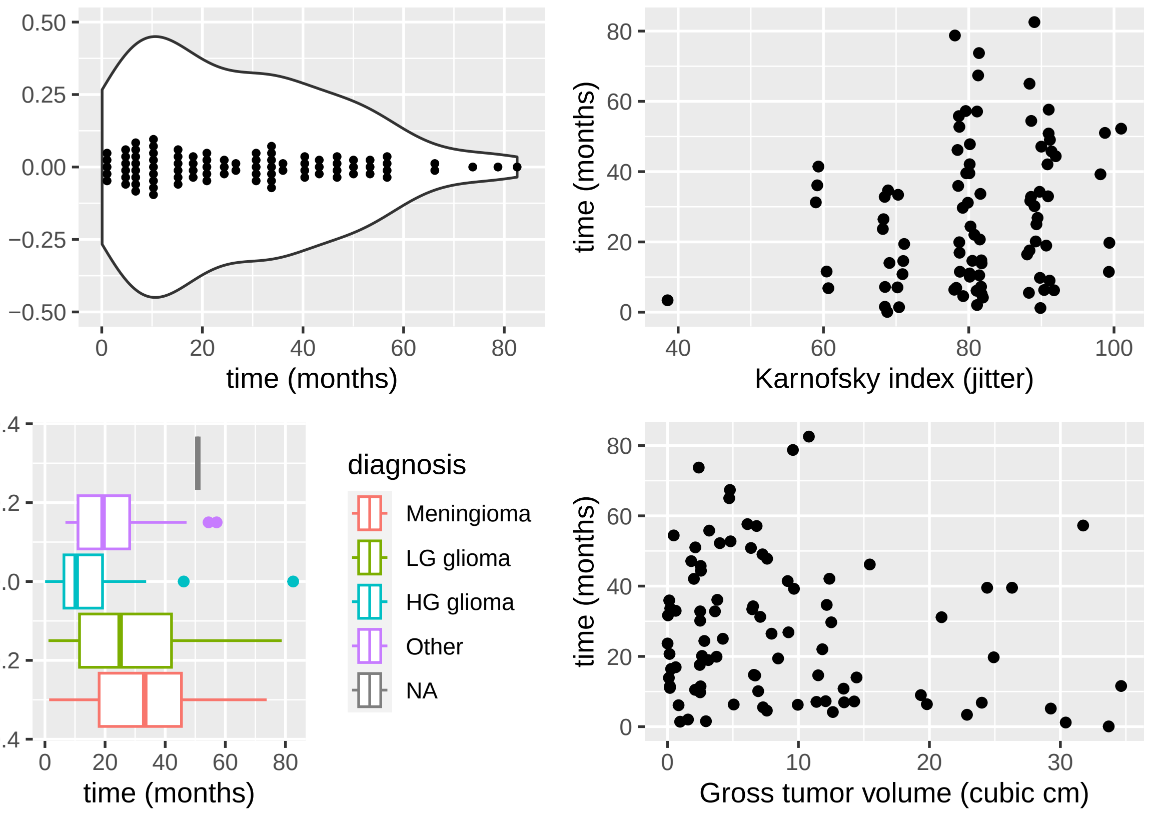
Suppose we have conducted a five-year medical study, in which the patients have been treated for cancer. We would like to fit a model to predict patient survival time, taking into account various features.

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).

```r
bc <- BrainCancer %>% mutate(status = as.factor(status))

vp <- bc %>% ggplot(aes(x=time, y=rep(0,88))) +
  geom_violin() +
  geom_dotplot(stackdir = "center", dotsize = 0.5) +
  labs(x = "time (months)", y = "")

kip <- bc %>% ggplot(aes(x = jitter(ki), y = time)) +
  geom_point() +
  labs(x = "Karnofsky index (jitter)", y = "time (months)")

dip <- bc %>% ggplot(aes(x = time, color = diagnosis)) +
  geom_boxplot() +
  labs(x = "time (months)")

gtvp <- bc %>% ggplot(aes(x = gtv, y = time)) +
  geom_point() +
  labs(x = "Gross tumor volume (cubic cm)", y = "time (months)")

grid.arrange(vp, kip, dip, gtvp, ncol=2)
```

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



1) What do we see in these figure?

a) we tend to see that, of those diagnosed with some sort of cancer, the majority of them do not survive; b) Those with a smaller Karnofsky index don't tend to live as long; c) The time of survival is shorter for those with HG glioma and higher with Meningioma; d) larger tumors tend to lead towards less survival times.

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

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

```
## Analysis of Variance Table
##
## Response: time
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## diagnosis  3  4105.2 1368.39  3.8905 0.01189 *
## ki         1  1913.2 1913.20  5.4395 0.02217 *
## gtv        1   172.6  172.64  0.4908 0.48556
## Residuals 81 28489.8  351.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

status: Whether the patient is still alive at the end of the study: 0=Yes, 1=No.

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

We will lose the information for those who survived past the observation time. But there is an important complication: hopefully some or many of the patients have survived until the end of the study. Such a patient's survival time is said to be censored: we know that it is at least five years, but we do not know its true value. We do not want to discard this subset of surviving patients, as the fact that they survived at least five years amounts to valuable information.

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:

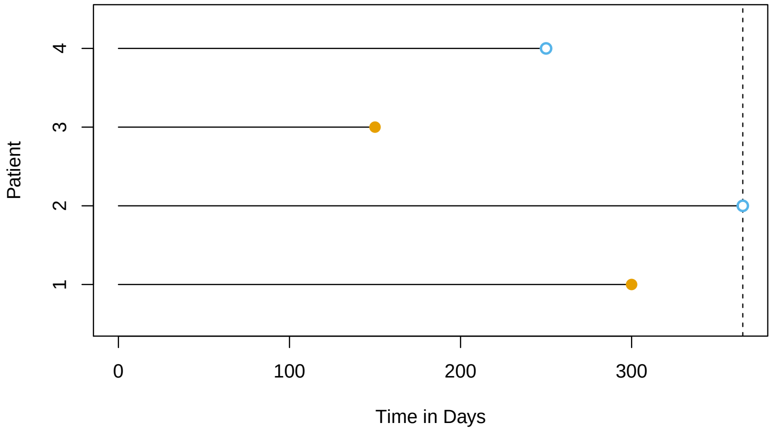
$$Y = \min(T, C) \tag{1}$$

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:

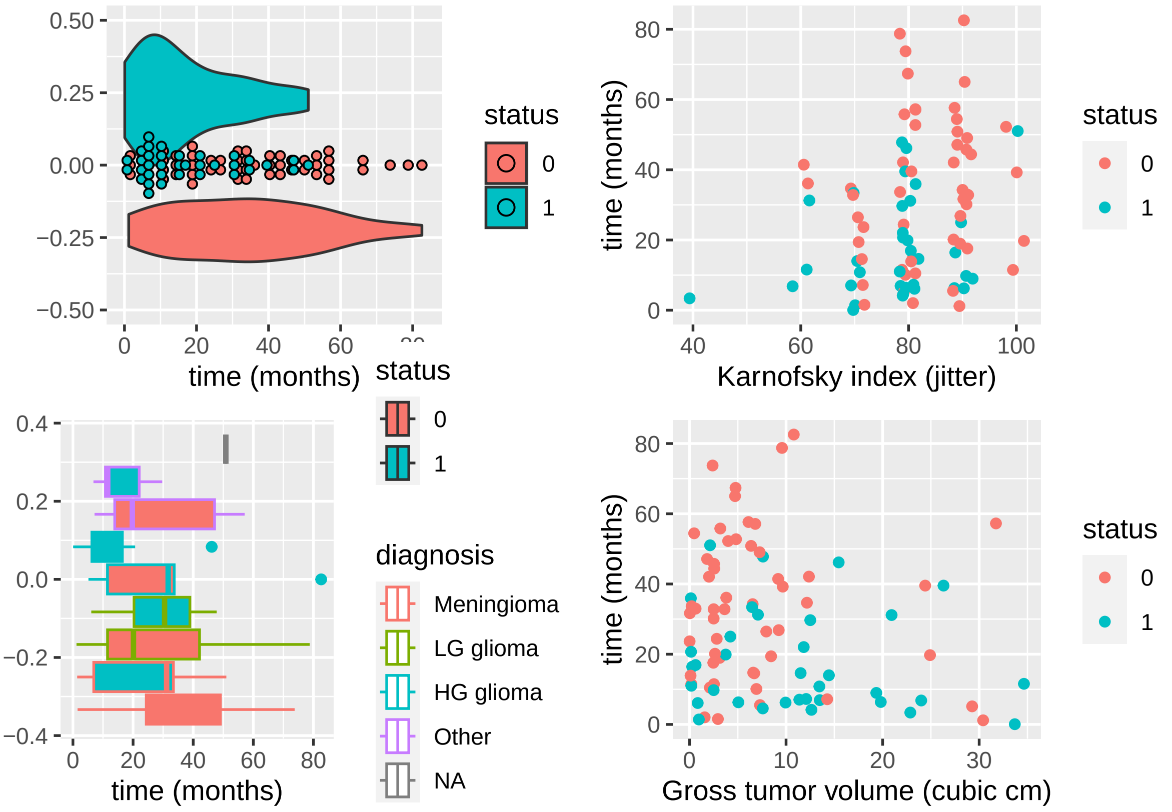


person 1: (300, 1); person 2: (365, 0); person 3: (150, 1); person 4: (250, 0). For patients 1 and 3, the event was observed. Patient 2 was alive when the study ended. Patient 4 dropped out of the study.

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

```r
vp2 <- bc %>% ggplot(aes(x=time, y=rep(0,88), fill=status)) +
  geom_violin() +
  geom_dotplot(aes(fill=status), stackdir = "center", dotsize = 0.95) +
  labs(x = "time (months)", y = "")

kip2 <- bc %>% ggplot(aes(x = jitter(ki), y = time, color=status)) +
  geom_point() +
  labs(x = "Karnofsky index (jitter)", y = "time (months)")

dip2 <- bc %>% ggplot(aes(x = time, color = diagnosis, fill=status)) +
  geom_boxplot() +
  labs(x = "time (months)")

gtvp2 <- bc %>% ggplot(aes(x = gtv, y = time, color=status)) +
  geom_point() +
  labs(x = "Gross tumor volume (cubic cm)", y = "time (months)")

grid.arrange(vp2, kip2, dip2, gtvp2, ncol=2)
```

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



words go here....

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 $F(t)$.

This decreasing function quantifies the probability of surviving past time t.
$P(T > t) = \int_t^\infty f(u)du = 1 - F(t)$

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?

```r
sum(bc$time > 20) / 88
```

```
## [1] 0.5454545
```

However, this does not seem quite right, since Y and T represent different quantities. In particular, 17 of the 40 patients who did not survive to 20 months were actually censored, and this analysis implicitly assumes that $T < 20$ for all of those censored patients; of course, we do not know whether that is true.

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 $t = 20$.

```r
# 71 not censored by time t=20
not.cens.by.t.20 <- 88 - sum(bc$status[bc$time <= 20] == 0)
# 48 survived past Y > 20
at.risk <- length(bc$time[bc$time > 20])

at.risk / not.cens.by.t.20
```

```
## [1] 0.6760563
```

However, this is not quite right either, since it amounts to completely ignoring the patients who were censored before time $t = 20$, even though the time at which they are censored is potentially informative. For instance, a patient who was censored at time $t = 19.9$ likely would have survived past $t = 20$ had he or she not been censored.

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

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

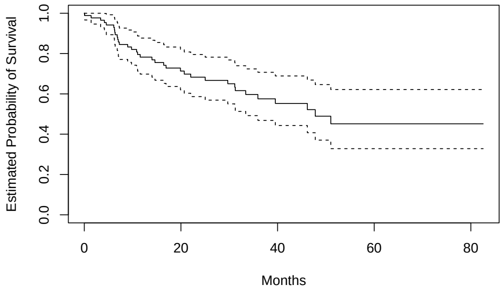
We now have the Kaplan-Meier estimator o the survival curve:

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

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?

```
# fit.surv2 <- survfit(Surv(time, status) ~ 1, data = bc)
# plot(fit.surv2 ,
#      xlab = "Months",
#      ylab = "Estimated Probability of Survival")

attach(BrainCancer)
fit.surv <- survfit(Surv(time, status) ~ 1)
plot(fit.surv, xlab = "Months",
    ylab = "Estimated Probability of Survival")
```



```
# get probability of survival for t = 20
ind <- tail(which(fit.surv$time < 20), n=1)
fit.surv$surv[ind]
```

```
## [1] 0.7131905
```

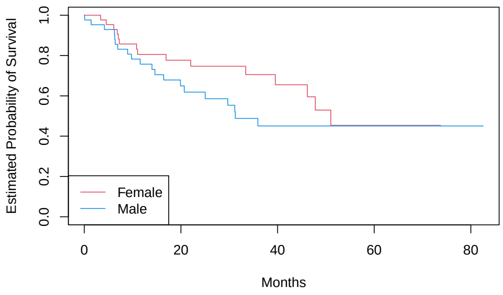
```
detach(BrainCancer)
```

Our estimate for $S(20) = 71.3\%$ which is larger that our two previous estimates that did not include the information from the censored data.

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.

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



```
# detach(BrainCancer)
```

The survival function for males tends to taper off quicker than females, but they then reamin (roughly) the same after about 50 months.

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?

we could test whether the mean survival time among the females equals the mean survival time among the males. But the presence of censoring again creates a complication since we don't know how long the censored values actually survived.

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(Surv(time, status) ~ sex)
logrank.test
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ sex)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=Female 45       15     18.5     0.676      1.44
## sex=Male   43       20     16.5     0.761      1.44
##
##  Chisq= 1.4  on 1 degrees of freedom, p= 0.2
```

The resulting p-value is 0.23, indicating no evidence of a difference in survival between the two sexes.

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:

$$\begin{aligned}
h(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t) \cap (T > t))/\Delta t}{P(T > t)} && \text{(by conditional prob.)} \\
&= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t))/\Delta t}{P(T > t)} && \text{(if } A \text{ then } B \text{ must be true)} \\
&= \frac{f(t)}{S(t)} && \text{(numerator is the PDF of } T \text{ and denominator is survival function)}
\end{aligned}$$

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:

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

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)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ sex)
##
##   n= 88, number of events= 35
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## sexMale 0.4077    1.5033   0.3420 1.192    0.233
##
##         exp(coef) exp(-coef) lower .95 upper .95
## sexMale     1.503     0.6652     0.769     2.939
##
## Concordance= 0.565  (se = 0.045 )
## Likelihood ratio test= 1.44  on 1 df,   p=0.2
## Wald test            = 1.42  on 1 df,   p=0.2
## Score (logrank) test = 1.44  on 1 df,   p=0.2
```

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
```

```
## [1] 1.440495
```

We see that the chisq value is very close to our LR/Wald test scores.... it is actually the same as the logrank test for a single explanatory variable

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(Surv(time, status) ~ sex + diagnosis + loc + ki + gtv + stereo)
summary(fit.all)
```

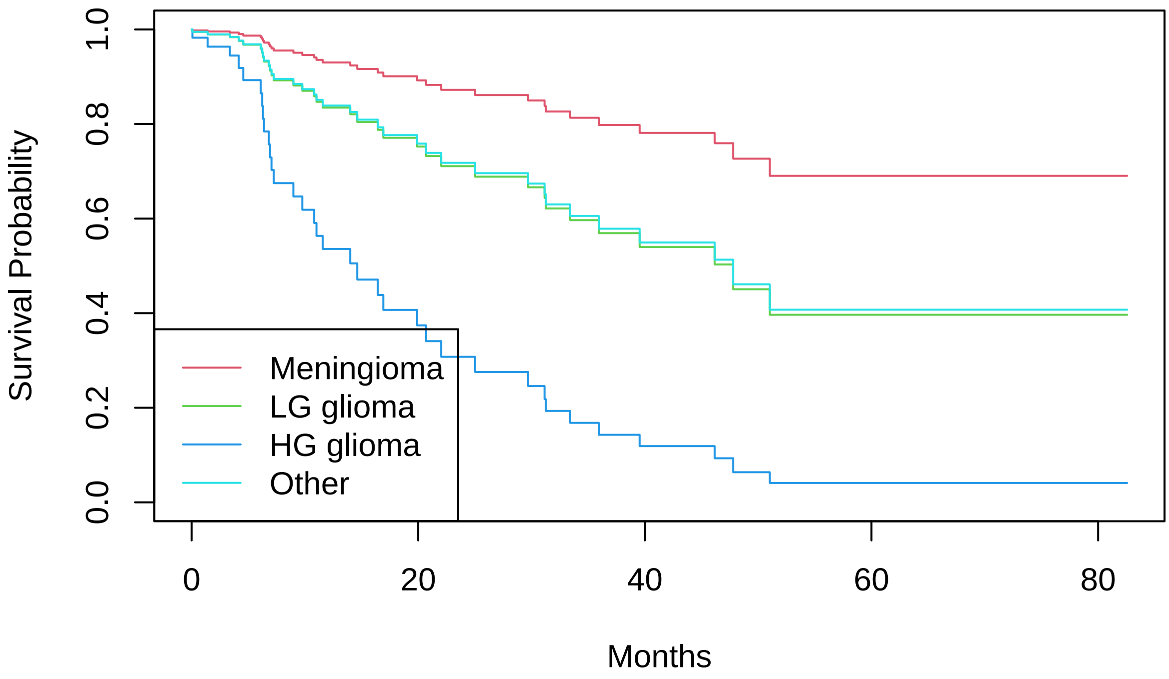
```
## Call:
## coxph(formula = Surv(time, status) ~ sex + diagnosis + loc +
##     ki + gtv + stereo)
##
##   n= 87, number of events= 35
##    (1 observation deleted due to missingness)
##
##                        coef exp(coef) se(coef)      z Pr(>|z|)
## sexMale             0.18375   1.20171  0.36036  0.510  0.61012
## diagnosisLG glioma  0.91502   2.49683  0.63816  1.434  0.15161
## diagnosisHG glioma  2.15457   8.62414  0.45052  4.782 1.73e-06 ***
## diagnosisOther      0.88570   2.42467  0.65787  1.346  0.17821
## locSupratentorial   0.44119   1.55456  0.70367  0.627  0.53066
## ki                 -0.05496   0.94653  0.01831 -3.001  0.00269 **
## gtv                 0.03429   1.03489  0.02233  1.536  0.12466
## stereoSRT           0.17778   1.19456  0.60158  0.296  0.76760
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## sexMale               1.2017     0.8321    0.5930    2.4352
## diagnosisLG glioma    2.4968     0.4005    0.7148    8.7215
## diagnosisHG glioma    8.6241     0.1160    3.5664   20.8546
## diagnosisOther        2.4247     0.4124    0.6678    8.8031
## locSupratentorial     1.5546     0.6433    0.3914    6.1741
## ki                    0.9465     1.0565    0.9132    0.9811
## gtv                   1.0349     0.9663    0.9906    1.0812
## stereoSRT             1.1946     0.8371    0.3674    3.8839
##
## Concordance= 0.794  (se = 0.04 )
## Likelihood ratio test= 41.37  on 8 df,   p=2e-06
## Wald test            = 38.7  on 8 df,   p=6e-06
## Score (logrank) test = 46.59  on 8 df,   p=2e-07
```

The diagnosis variable has been coded so that the baseline corresponds to meningioma. The results indicate that the risk associated with HG glioma is more than eight times (i.e. $e^2.15 = 8.62$) the risk associated with meningioma. In other words, after adjusting for the other predictors, patients with HG glioma have much worse survival compared to those with meningioma. In addition, larger values of the Karnofsky index, ki, are associated with lower risk, i.e. longer survival.

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)
```


Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
Months
Meningioma
LG glioma
HG glioma
Other