Survival Analysis and Censored Data Exercises

Simone Collier

The Kaplan-Meier Survival Curve

Start by loading the packages we need. If you need to install the packages first then run `install.packages("PACKAGENAME")` in your console before running the code chunk.

```
library(ISLR2)
library(survival)
```

We will be making use of the `BrainCancer` data set in the `ISLR2` package. It contains the survival times for patients with primary brain tumors undergoing treatment. It also contains information for serveral predictor variables which we will discuss later. 53 of the 88 patients were still alive at the end of the study.

```
attach(BrainCancer)
```

We should first check how the `status` variables (δ) has been coded.

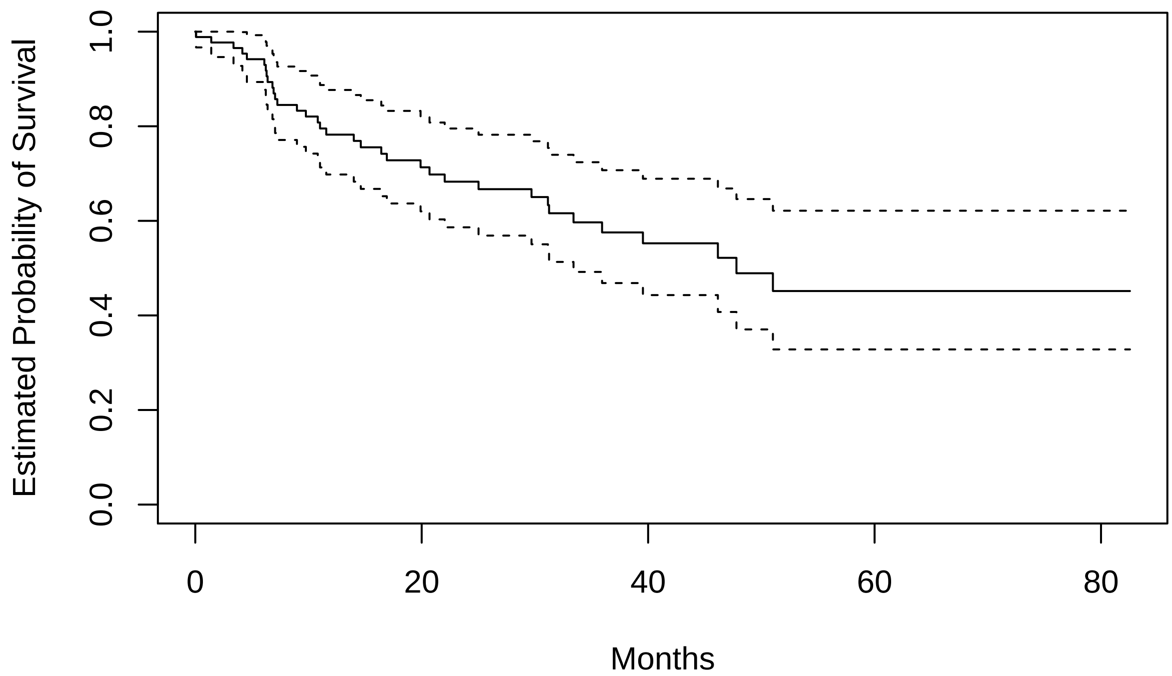
```
table(status)
```

```
## status
##  0  1
## 53 35
```

We see there are 53 patients with $status = 0$ so this means `status = 0` means the survival time is censored and `status = 1` means the survival time is uncensored.

We can create the Kaplan-Meier survival curve using the `survfit()` function within the `survival` library. Note that `time` is y_i, the time to the i-th event (censoring or death).

```
survival.curve <- survfit(Surv(time, status) ~ 1)
plot(survival.curve, xlab = "Months", ylab = "Estimated Probability of Survival")
```

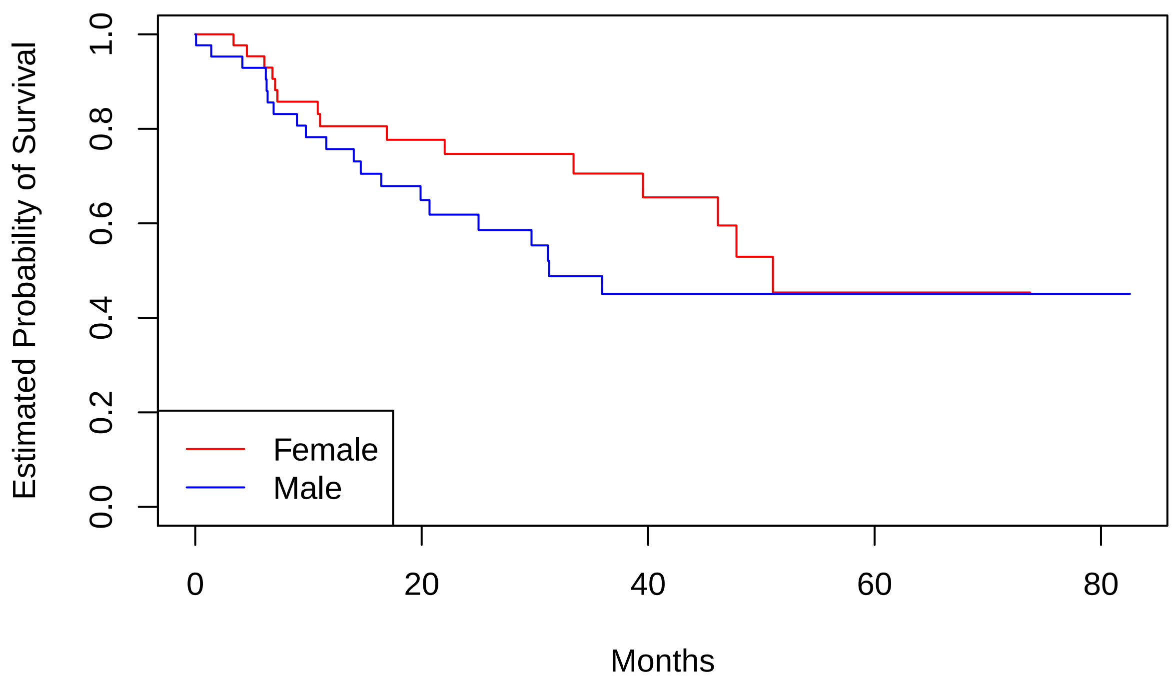


The solid line is the Kaplan-Meier survival curve and the dashed lineas are the standard error bands.

The Log-Rank Test

Now we want to create Kaplan-Meier survival curves from the `BrainCancer` data that are separated by sex.

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



We can use a log-rank test to compare the survival curves for males versus females. We use the `survdiff()` function from the `survival` package.

```
logrank <- survdiff(Surv(time, status) ~ sex)
logrank
```

```
## 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 0.2 which is greater than 0.05, so there is no evidence of a difference in the survival between the two sexes.

The Cox Proportional Hazards Model

We have seen the restults from the log-rank test that compares the survival curves for males versus females form the `BrainCancer` data. Now we can fit a Cox proportional hazards model to test the exact samme thing. We use the `coxph()` function from the `survival` library.

```
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
```

The p-value for the hypothesis test $H_0 : \beta = 0$ is 0.233 which is not significant so we conclude that there is no difference in the survival rates between males and females. This is the same conclusion we found for the log-rank test.

Now let's try to fit a model with multiple predictors. Note that the covariates included are either quanititative or qualitative with a binary response with the exception of `diagnosis`. The `diagnosis` variable has four classes: `Meningioma`, `LG glioma`, `HG glioma`, and `Other`. The `coxph()` function automatically chooses the first class as the baseline for all qualitative variables.

```r
multi.fit.cox <- coxph(Surv(time, status) ~ sex + diagnosis + loc + ki + gtv + stereo)
multi.fit.cox
```

```
## Call:
## coxph(formula = Surv(time, status) ~ sex + diagnosis + loc +
##     ki + gtv + stereo)
##
##                         coef exp(coef) se(coef)      z        p
## 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
##
## Likelihood ratio test=41.37  on 8 df, p=1.776e-06
## n= 87, number of events= 35
##    (1 observation deleted due to missingness)
```

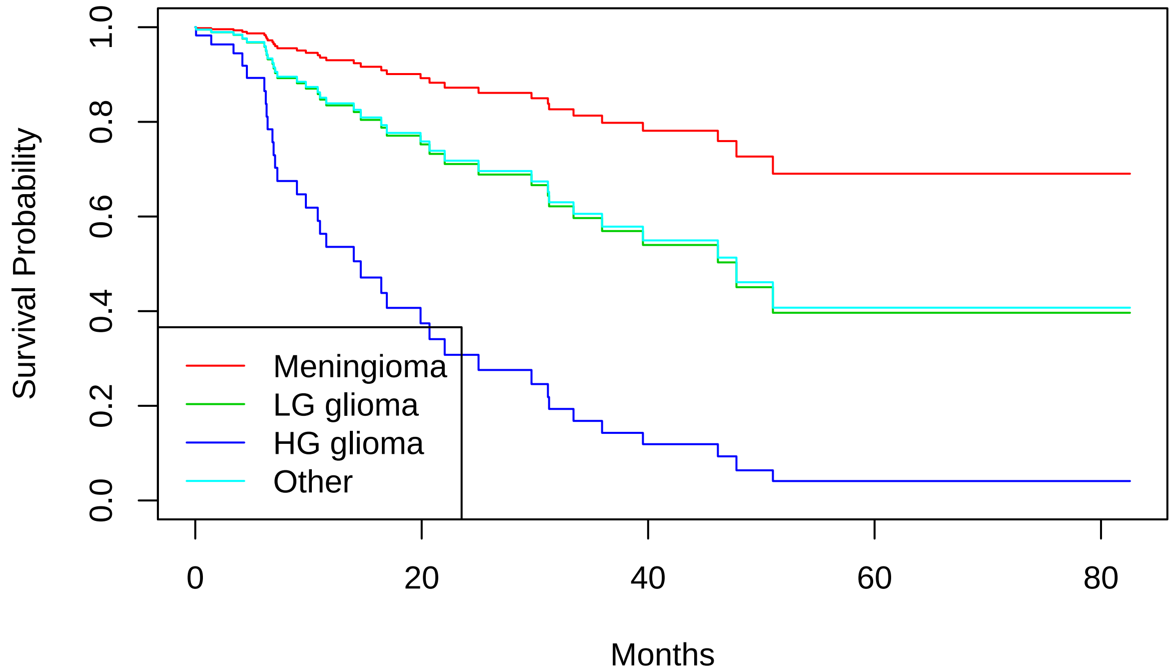
Since `Meningioma` was coded as the baseline, the fitted coefficient 2.15 associated with `HG glioma` means that the risk associated with `HG glioma` is $e^{2.15} = 8.62$ times more that the risk of `Meningioma`.

We can plot the survival curves for each diagnosis category while adusting for the other predictors. To make these plots we make a new data set where the value for each of the other predictors is the mean (if quantitative) or mode (if qualitative) of the variable. So,

```r
plot.data <- 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))
```

Now we use the `survfit()` function with our fitted model and `plot.data` as the `newdata`.

```r
survplots <- survfit(multi.fit.cox, newdata = plot.data)
plot(survplots, xlab = "Months", ylab = "Survival Probability", col = 2:5)
legend("bottomleft", levels(diagnosis), col = 2:5, lty = 1)
```



These exercises were adapted from : James, Gareth, et al. An Introduction to Statistical Learning: with Applications in R, 2nd ed., Springer, 2021.