BIOS522: Survival Analysis Methods

Activity 3: R Practice Session

In this session, you will practice implementing basic analyses in `R`. You will use `R markdown` to document your code and results. `R` chunks are provided for your code, and space is provided for your interpretations.

For this and other `R` practice sessions, we will use the data set `whas500`. This data set includes data extracted from the Worcester Heart Attack Study (WHAS). The study is described in Hosmer, Lemeshow and May (2013). The purpose of the study is to study factors and time trends associated with long-term survival following acute myocardial infarction. The main data set has information on more than 11,000 admissions. The data in `whas500` were sampled by taking an approximately 23% random sample from cohorts collected in 1997, 1999, and 2001, yielding 500 subjects.

The file `whas500.Rdata` includes the data. The file `whas500.txt` summarizes the variables.

1. Use the code chunk below to load the `survival` package, load the `whas500` data set (if you aren't sure how, figuring it out is good practice!), and inspect its variables. Refer to the documentation to review the meaning of each variable.

```
# ADD YOUR CODE HERE
library(survival)
setwd("C:/Users/nedean/OneDrive - Emory University/Teaching/BIOS522/R coding")
load("whas500.RData")
whas500[1:10,]
```

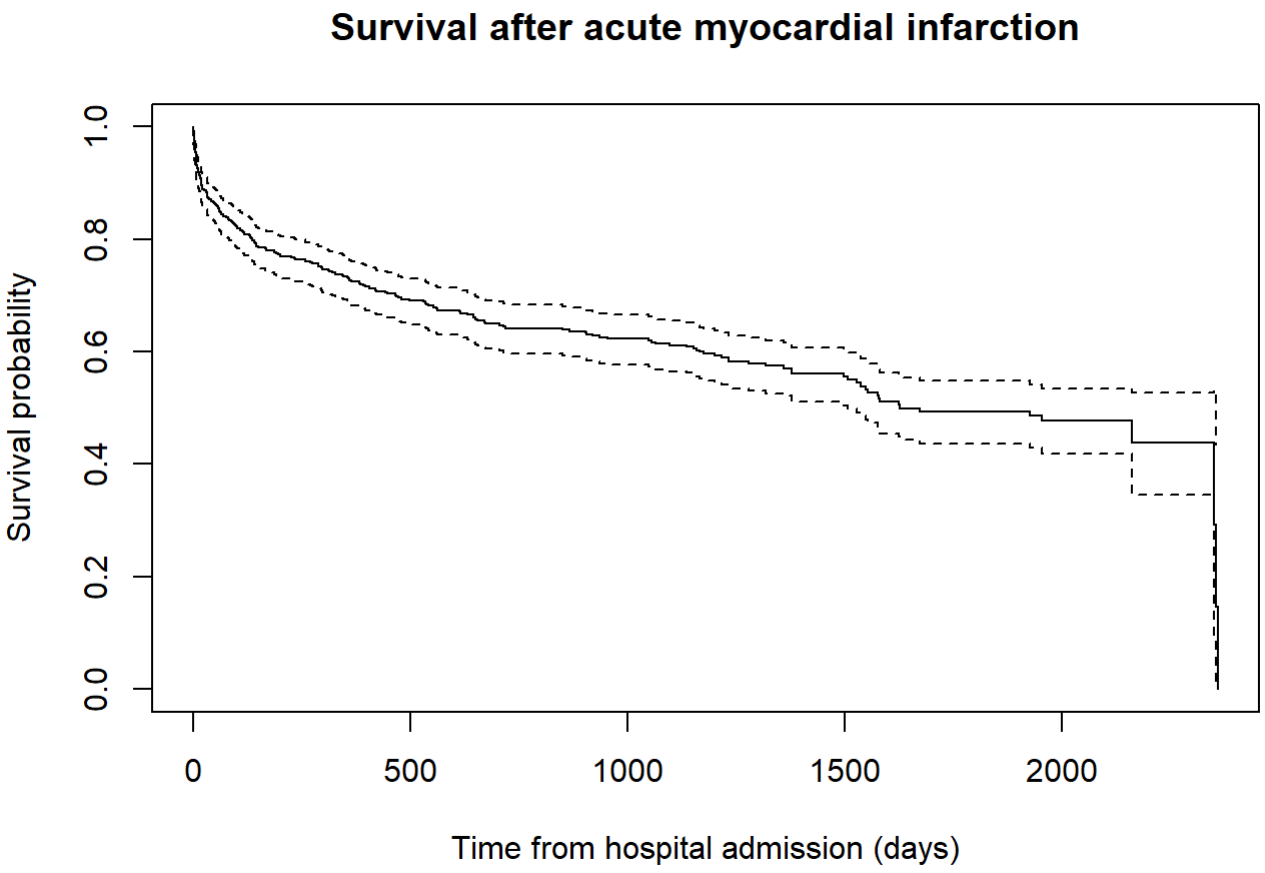
```
##    id age gender  hr sysbp diasbp      bmi cvd afb sho chf av3 miord mitype
## 1   1  83      0  89   152     78 25.54051   1   1   0   0   0     1      0
## 2   2  49      0  84   120     60 24.02398   1   0   0   0   0     0      1
## 3   3  70      1  83   147     88 22.14290   0   0   0   0   0     0      1
## 4   4  70      0  65   123     76 26.63187   1   0   0   1   0     0      1
## 5   5  70      0  63   135     85 24.41255   1   0   0   0   0     0      1
## 6   6  70      0  76    83     54 23.24236   1   0   0   0   1     0      0
## 7   7  57      0  73   191    116 39.49046   1   0   0   0   0     0      1
## 8   8  55      0  91   147     95 27.11609   1   0   0   0   0     0      1
## 9   9  88      1  63   209    100 27.43554   1   0   0   1   0     0      0
## 10 10  54      0 104   166    106 25.54448   1   0   0   0   0     0      0
##    year  admitdate    disdate      fdate los dstat lenfol fstat
## 1     1 01/13/1997 01/18/1997 12/31/2002   5     0   2178     0
## 2     1 01/19/1997 01/24/1997 12/31/2002   5     0   2172     0
## 3     1 01/01/1997 01/06/1997 12/31/2002   5     0   2190     0
## 4     1 02/17/1997 02/27/1997 12/11/1997  10     0    297     1
## 5     1 03/01/1997 03/07/1997 12/31/2002   6     0   2131     0
## 6     1 03/11/1997 03/12/1997 03/12/1997   1     1      1     1
## 7     1 03/10/1997 03/15/1997 12/31/2002   5     0   2122     0
## 8     1 01/11/1997 01/15/1997 02/15/2001   4     0   1496     1
## 9     1 12/31/1996 01/04/1997 07/09/1999   4     0    920     1
## 10    1 01/16/1997 01/21/1997 12/31/2002   5     0   2175     0
```

2. How is `lenfol` calculated? What is the time origin?

`lenfol` is the difference between admitdate (hospital admission date) and fdate (date of last-follow-up). The time origin is hospital admission date for acute myocardial infarction.

3. Plot a Kaplan-Meier curve for survival after acute myocardial infarction. Use the entire dataset. Plot the curve with complementary log-log confidence intervals. Make sure the plot has proper labels.

```
# ADD YOUR CODE HERE
whasKM <- survfit(Surv(lenfol,fstat) ~ 1, data=whas500, conf.type="log-log")
plot(whasKM,xlab = "Time from hospital admission (days)", ylab = "Survival probability",
     main = "Survival after acute myocardial infarction")
```



4. Comment on the shape of the survival curve. How does survival change over time? What happens to the standard errors over time?

There is a sharp decrease early on, capturing deaths occurring shortly after the acute event. Eventually the rate of death slows, though steadily declines over time. The standard errors are increasing over time. The standard error increases as the number of people at risk decreases.

5. Calculate the estimated survival at 1 and 2 years. Report point estimates and 95% complementary log-log confidence intervals.

```
# ADD YOUR CODE HERE
summary(whasKM,times=c(365,730))
```

```
## Call: survfit(formula = Surv(lenfol, fstat) ~ 1, data = whas500, conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   365    362     138    0.724  0.0200        0.683        0.761
##   730    236      34    0.642  0.0222        0.596        0.683
```

In this population, estimated survival one year after acute myocardial infarction is 72.4% (95% confidence interval 68.3%, 76.1%). Esimated survival two years after acute myocardial infarction is 64.2% (95% confidence interval 59.6%, 68.3%).

6. If possible, calculate median survival and an accompanying 95% confidence interval. Include units.

```
# ADD YOUR CODE HERE
whasKM
```

```
## Call: survfit(formula = Surv(lenfol, fstat) ~ 1, data = whas500, conf.type = "log-log")
##
##        n events median 0.95LCL 0.95UCL
## [1,] 500    215   1627    1506    2353
```

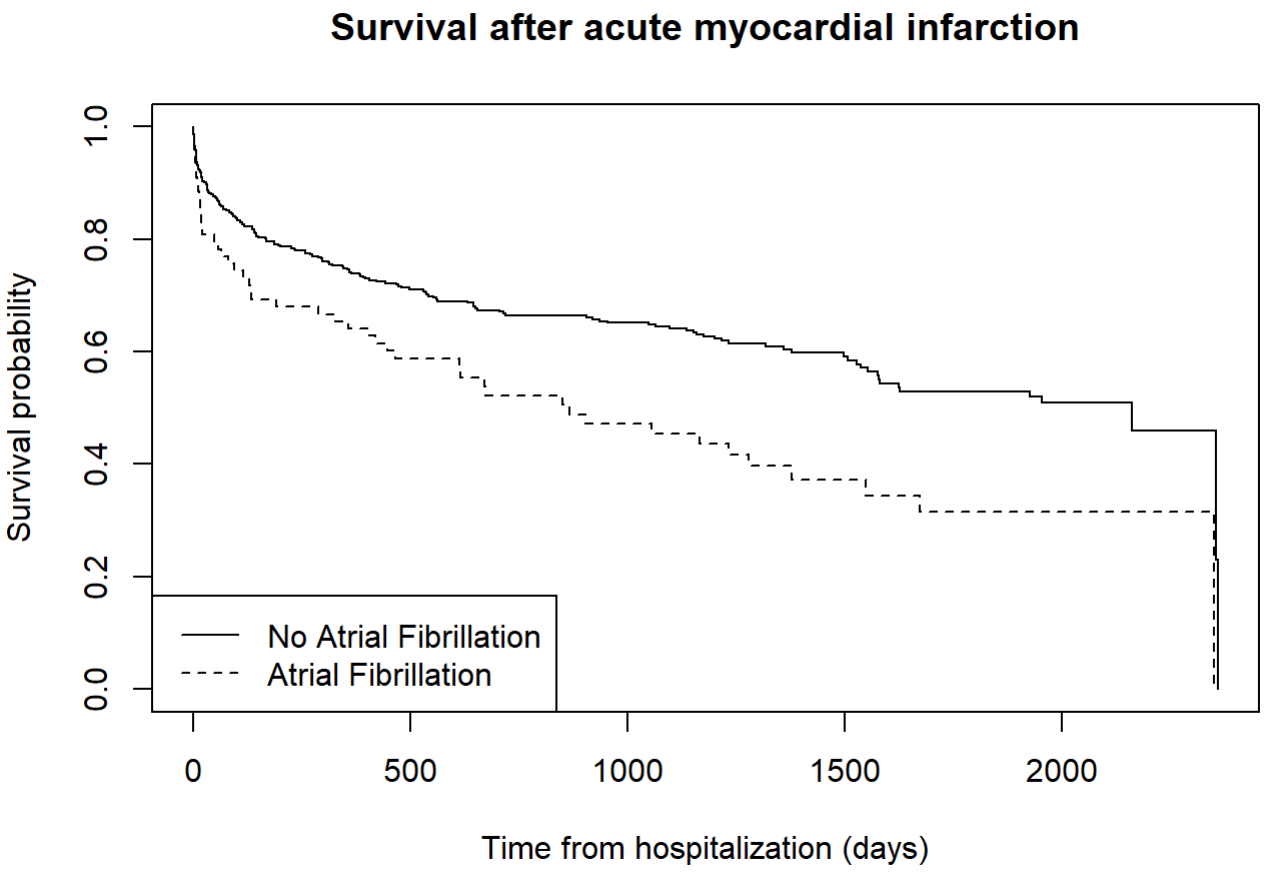
In this population, estimated median survival after acute myocardial infarction is 1627 days (95% confidence interval 1506, 2363 days).

7. Whether the patient required atrial fibrillation during the acute myocardial infarction is a potentially important predictor of survival. Plot separate Kaplan-Meier curves by atrial fibrillation status. Include a legend. Comment on any observable differences.

```
# ADD YOUR CODE HERE
whasKM.afb <- survfit(Surv(lenfol,fstat) ~ afb, data=whas500)
whasKM.afb
```

```
## Call: survfit(formula = Surv(lenfol, fstat) ~ afb, data = whas500)
##
##         n events median 0.95LCL 0.95UCL
## afb=0 422    168   2160    1577      NA
## afb=1  78     47    865     465    1548
```

```
plot(whasKM.afb, lty=c(1,2),
     xlab = "Time from hospitalization (days)", ylab = "Survival probability",
     main = "Survival after acute myocardial infarction")
legend("bottomleft", legend=c("No Atrial Fibrillation","Atrial Fibrillation"),
      lty=c(1,2))
```



Patients requiring atrial fibrillation had much poorer survival than patients who did not require atrial fibrillation. One can see this in separation of the survival curves and also in the median survival (2160 days for those without atrial fibrillation vs. 865 days for those with atrial fibrillation).

8. Calculate a log-rank test statistic to test the statistical significance of the difference in survival between patients with atrial fibrillation and with no atrial fibrillation.

```
# ADD YOUR CODE HERE
survdiff(Surv(lenfol,fstat) ~ afb, data=whas500)
```

```
## Call:
## survdiff(formula = Surv(lenfol, fstat) ~ afb, data = whas500)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## afb=0 422      168    184.8      1.52      10.9
## afb=1  78       47     30.2      9.31      10.9
##
##  Chisq= 10.9  on 1 degrees of freedom, p= 0.001
```

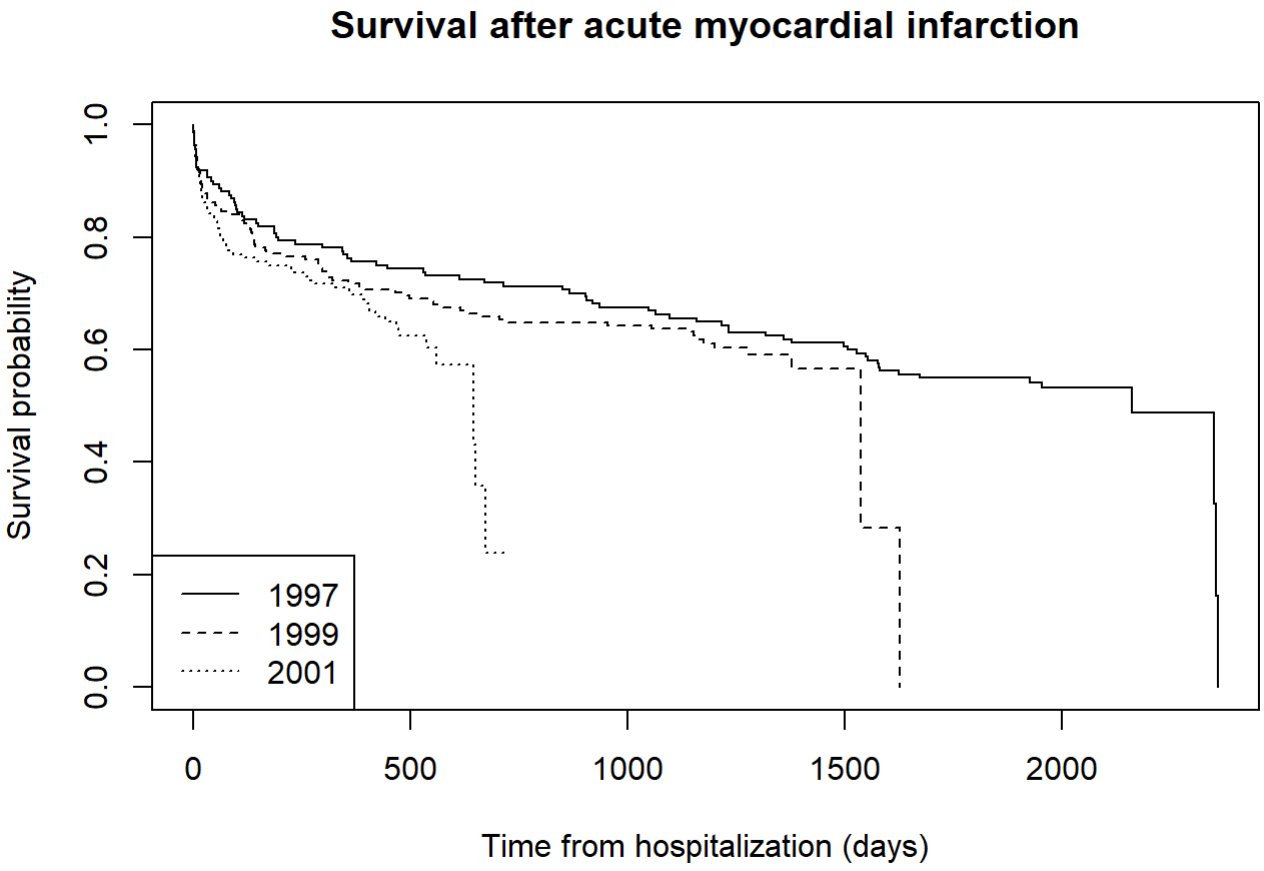
Patients requiring atrial fibrillation had statistically significantly poorer survival than patients not requiring atrial fibrillation (log-rank test chi-squared test statistic = 10.9, p-value = 0.001).

9. Patients were enrolled from three different cohort years - 1997, 1999, and 2001. Plot separate Kaplan-Meier curves by cohort. Include a legend. Comment on any observable differences.

```
# ADD YOUR CODE HERE
whasKM.yr <- survfit(Surv(lenfol,fstat) ~ year, data=whas500)
whasKM.yr
```

```
## Call: survfit(formula = Surv(lenfol, fstat) ~ year, data = whas500)
##
##          n events median 0.95LCL 0.95UCL
## year=1 160     78   2160    1577      NA
## year=2 188     77   1536    1377      NA
## year=3 152     60    646     559      NA
```

```
plot(whasKM.yr, lty=c(1,2,3),
     xlab = "Time from hospitalization (days)", ylab = "Survival probability",
     main = "Survival after acute myocardial infarction")
legend("bottomleft", legend=c("1997","1999","2001"),
      lty=c(1,2,3))
```



Survival across the three cohorts is similar, although slightly worse for later cohorts. Note the differences in follow-up time across the three groups. As we would expect, patients enrolled in 1997 are followed for longer than patients from 1999 or from 2001.

10. Recalculate the log-rank test for atrial fibrillation, adjusting for cohort.

```
# ADD YOUR CODE HERE
survdiff(Surv(lenfol,fstat) ~ afb + strata(year), data=whas500)
```

```
## Call:
## survdiff(formula = Surv(lenfol, fstat) ~ afb + strata(year),
##     data = whas500)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## afb=0 422      168    184.1      1.41        10
## afb=1  78       47     30.9      8.38        10
##
##  Chisq= 10  on 1 degrees of freedom, p= 0.002
```

Adjusting for cohort using a stratified log-rank test has little impact on the final conclusion. Patients requiring atrial fibrillation had statistically significantly poorer survival than patients not requiring atrial fibrillation, adjusting for cohort year (stratified log-rank test chi-squared test statistic = 10, p-value = 0.002).

11. Imagine that we wanted to adjust our atrial fibrillation comparison for both cohort year and patient age at hospital admission. Describe how we might achieve this…

We could create a categorical covariate for patient age (e.g. below or above some cutoff) and then create a new categorical variable that considers all combinations of cohort and age category (i.e. 1997 younger, 1997 older, 1999 younger, 1999 older, 2001 younger, 2001 older). We could then conduct a stratified log-rank test. Alternatively, we could use a regression model framework that allows us to adjust for continuous covariates, or to adjust for multiple covariates without creating a separate group for all possible combinations… stay tuned to learn more about this!

End of practice session