This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the Knit button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(survival)

## Loading required package: splines

head(heart)
```

<table>
<thead>
<tr>
<th></th>
<th>start</th>
<th>stop</th>
<th>event</th>
<th>age</th>
<th>year</th>
<th>surgery</th>
<th>transplant</th>
<th>id</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>50</td>
<td>1</td>
<td>-17.155373</td>
<td>0.1232033</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>3.835729</td>
<td>0.2546201</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6.297057</td>
<td>0.2655715</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>16</td>
<td>1</td>
<td>6.297057</td>
<td>0.2655715</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>36</td>
<td>0</td>
<td>-7.737166</td>
<td>0.4900753</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>36</td>
<td>39</td>
<td>1</td>
<td>-7.737166</td>
<td>0.4900753</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

What is the event in this dataset? It’s labeled as “event” and marks the death of a cardiac patient tracked in this study.

Are there treatments to be considered in this dataset? “Transplant” indicates whether or not the patient received a transplant.

Are there other covariates and what are they? “Age” indicates the age of the patient (-47). “Year” indicates year since 1967. “Surgery” indicates whether the patient has previously had surgery.

How can we determine the time-to-event? Should be “stop”–“start”. These are measured in days

Is the data censored? The data is somewhat complicated in that patients come in to the study at different times presumably following some sort of heart trouble. Some leave the study by dying, others appear to live past the end of the study, which we can determine by calculating the following from the “stop” and “start” columns

```
max(heart$stop-heart$start)/365
```

```
## [1] 4.863014
```

as 4.86 years. So the data is right-censored.

We can now fit the Cox proportional hazards model. Turns out this function can handle multiple events per subject! Notice we can also just include the start-stop arithmetic in the formula!

First, though, let’s test that the assumptions of the Cox regression model fit are met. These are:

1. Non-informative censoring - the underlying study must ensure that the mechanisms giving rise to censoring of individual subjects are not related to the probability of an event occurring. For example, in clinical studies, care must be taken that continuation of follow-up not depend on a participants medical condition. Violation of this assumption can invalidates just about any sort of survival analysis, from Kaplan-Meier estimation to the Cox model.
The second key assumption in the Cox model is proportionality of hazards. In regression this means that the survival curves for two strata (determined by the particular choices of values for the x-variables) must have hazard functions that are proportional over time (i.e. constant relative hazard).

```r
fit<-coxph(formula=Surv((stop-start),event)~ year + age + surgery + cluster(id) + strata(transplant), data=heart)
test<-cox.zph(fit)
```

In the output, none of the covariates are significant (p<0.05) indicating that the Cox proportional hazards assumptions hold.

Now let’s look at a summary of the model.

```r
cph.heart<-coxph(formula=Surv((stop-start),event)~ year + age + surgery + cluster(id) + strata(transplant), data=heart)
summary(cph.heart)
```

```
## Call:
## coxph(formula = Surv((stop - start), event) ~ year + age + surgery +
##     cluster(id) + strata(transplant), data = heart)
##
## n= 172, number of events= 75
##
## coef exp(coef) se(coef) robust se z Pr(>|z|)    
##  year  -0.15745  0.85432  0.07087   0.07302 -2.156  0.0311 *   
##  age   0.02843  1.02883  0.01394   0.01344  2.116  0.0344 *   
## surgery -0.65143  0.52130  0.36910   0.36797 -1.770  0.0767 .    
##
## exp(coef) exp(-coef) lower .95 upper .95
##  year  0.8543  1.171  0.7404  0.9858          
##  age   1.0288  0.972  1.0021  1.0563          
## surgery 0.5213  1.918  0.2534  1.0723          
##
## Concordance= 0.637  (se = 0.055 )
## Rsquare= 0.09  (max possible= 0.958 )
## Likelihood ratio test= 16.27 on 3 df,  p=0.0009976
## Wald test = 17.4  on 3 df,  p=0.0005855
## Score (logrank) test = 16.11 on 3 df,  p=0.001075,  Robust = 18.42  p=0.0003603
##
## (Note: the likelihood ratio and score tests assume independence of
##  observations within a cluster, the Wald and robust score tests do not).
```

How do we interpret these results?

“Year” appears to have a significantly negative effect on mortality. Remember that year is the years since November 1967 that a patient was accepted to the transplant program. So mortality decreases with time presumably because of technological or scientific progress.

“Age” is positively associated with mortality. No surprises there.

“Surgery”, previous bypass surgery, is negatively related to mortality, but only marginally significant.

How to plot? First, we fit time curves by strata. Basically, we’re predicting, using the regression model across the matrix of covariates.
What are the interesting features of this graph?