Survival Models

Benjamin Schneider

Table of contents

Motivation and Overview

Why survival models?

- Survival models are useful when your output variable surrounds a failure event over a period of time. These have been traditionally used as the following:
  - Survival of patients under certain conditions
  - Failure of mechanical equipment
  - Disease occurance or reoccurance
- Political scientists can also use these ideas in several different settings.

Preview of models covered

- Survival models have several different available methods for estimation. Today we will focus on:
  - Log-rank tests
  - Cox proportional hazards regression

Introduction to Survival Models

Conceptualizing Survival

- When working with data we are modeling with survival analysis, we consider the survival function (Note this is the complement of the CDF):

$$S(t) = \mathrm{P}\{T \geq t\} = 1 - F(t) = \int_t^{\infty} f(x)dx$$

- Where $f(x)$ is the PDF.
- Another conception is the hazard function which can be thought of as the intantaneous rate of your event of interest:

$$\lambda(t) = \lim_{dt \to 0} \frac{\mathrm{P}\{t \leq T < t + dt | T \geq t\}}{dt}$$

- It is useful to think this way often in political science modeling as most of our events we are modeling will have nonzero probabilities at all times.

Conceptualizing Survival

- Using a little bit of algebra, we can rewrite the Hazard function as:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

- This can be an alternative way to think about it where the hazard function is the pdf over the survival function.

Censoring

- Censoring is a form of missing data issues with survival models
- Right censoring occurs when you have information on the lower limit but not the upper limit of one of your cases (i.e. a patient does not follow up)
- Left censoring occurs when the lower limit is unknown. This is commonly the case with managing diseases and not knowing the time of onset.
- Right censoring is much easier to deal with in the models we talk about today, though techniques exist for left censoring.

Log-rank tests

Log-rank tests overview

- Log rank tests are used to compare the survival conditions of two distinct groupings.
- This is commonly a treatment and control group, but it can be any pariwise comparison you may want to get an initial estimate of differences about.
- Think of this as the difference-in-means of survival models

Log-rank tests overview

- The log-rank test is the most commonly used test in survival analysis
- The statistic compares the hazard functions of two different groups at the same time.
- Consider that for each time j the number of subjects in each are N_{1j} and N_{2j}. Where: $N_{1j} + N_{2j} = N_j$
- Let the number of observed events (failure, death) for each of the groups at time j be prepresented as O_{1j} and O_{2j}. Where: $O_{1j} + O_{2j} = O_j$

Log-rank tests overview

- The null hypotheses (both groups are identical) is modeled by $O_{1j} \sim \textit{hypergeometric}(N_j, N_{1j}, O_j)$
- The distribution has an expectation of E_{1j} and a variance of V_j
- The log rank statistic compares O_{1j} with its expectation from the distribution and is represented as:

$$Z = \frac{\sum_{j=1}^{J}(O_{1j} - E_{1j})}{\sum_{j=1}^{J} V_j}$$

Estimating the log-rank test

- Some useful packages to use when estimating and visualizing surival models are `survival` and `survminer` which uses the visualization package `ggplot2`.

```
library(survival)
library(survminer)
set.seed(213)
control<-cbind(1,round(rweibull(100,shape=1)*100))
treatment<-cbind(2,round(rweibull(100,shape=3)*100))
```

Estimating the log-rank test

```
dataset<-rbind(control,treatment)
colnames(dataset)<-c('group','time')
living<-c()
for(i in 1:nrow(dataset)){
  if(dataset[i,2]>=75){
    living<-c(living,F)
    dataset[i,2]=75
  }
  else{
    living<-c(living,T)
  }
}
dataset<-cbind(dataset,living)
dataset<-as.data.frame(dataset)
```

Estimating the log-rank test

```
survivalobject<-Surv(time=dataset$time,event=dataset$living)
model<-survfit(survivalobject~dataset$group)
survdiff(survivalobject~dataset$group)
```

```
## Call:
## survdiff(formula = survivalobject ~ dataset$group)
##
##                      N Observed Expected (O-E)^2/E (O-E)^2/V
## dataset$group=1 100       53     36.1      7.88      13.8
## dataset$group=2 100       32     48.9      5.82      13.8
##
##  Chisq= 13.8  on 1 degrees of freedom, p= 0.000199
```

- At the time I truncated the model ($t = 75$) we can see that the difference between the difference between the observed and expected events for both groups.

Estimating the log-rank test

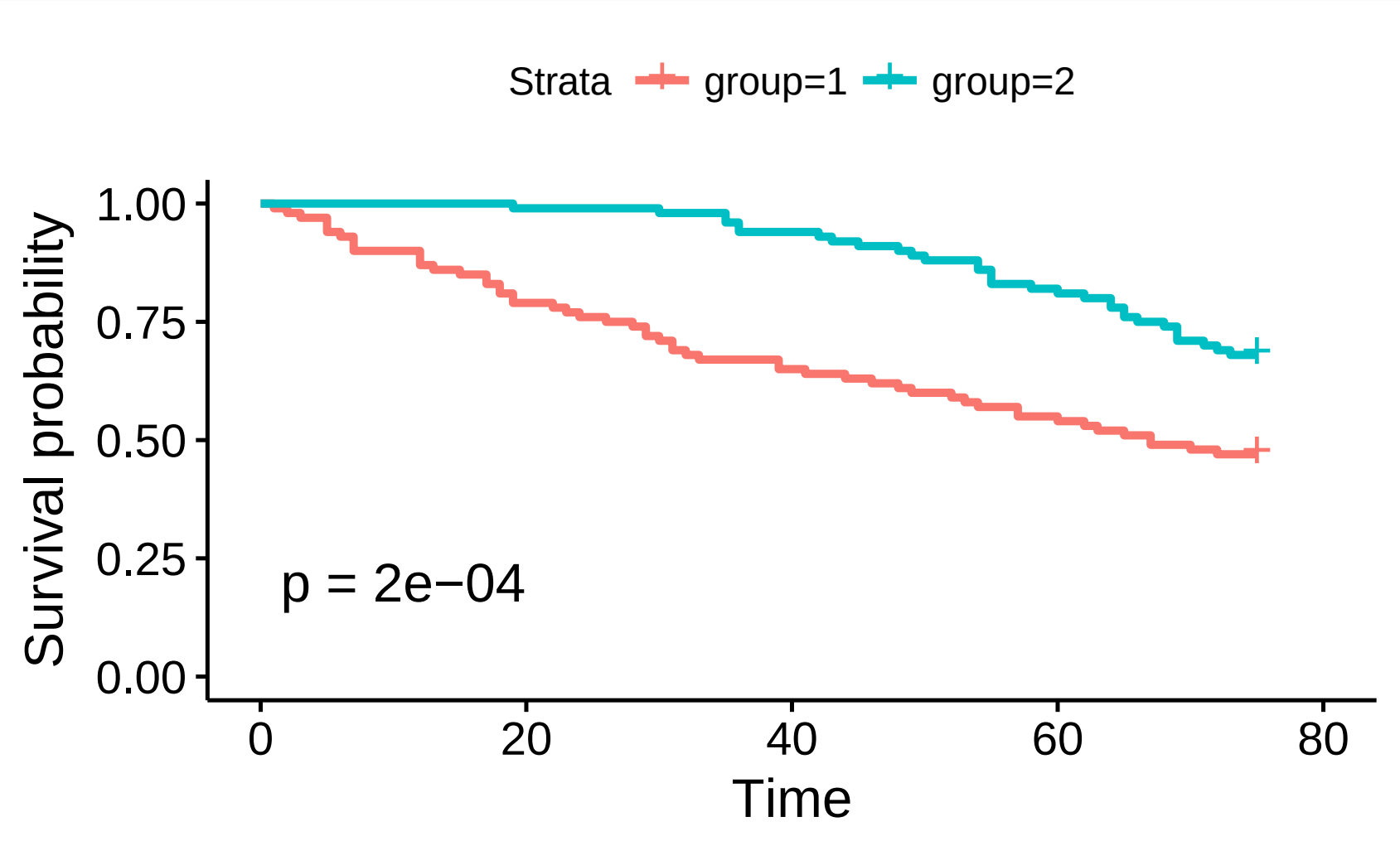
```
library(coin)
logrank_test(survivalobject~as.factor(dataset$group))

##
##  Asymptotic Two-Sample Logrank Test
##
## data:  survivalobject by as.factor(dataset$group) (1, 2)
## Z = -3.674, p-value = 0.0002388
## alternative hypothesis: true theta is not equal to 1
```

- I use the logrank test function from the `coin` library to output the simple log-rank test output.
- This function evaluates the statistic and shows that there is a significant difference between the groups.
- Survival models are usually also interpreted visually

Estimating the log-rank test

```
ggsurvplot(model, data = dataset, pval = TRUE)
```



Estimating the log-rank test

- As seen in the previous slide, the survminer and ggplot libraries work together well to visualize our two different groups using a Kaplan-Meier curve.
- We can see here that the patients in group 2 did much better.
- We can also assess general trends. For instance, we can see that the treatment group performs very well early on and then they even out.

Cox proportional hazard models

Cox proportional hazards model overview

- When you want to assess multiple covariates at once, the Cox proportional hazards model will be the ideal model.
- In order to use this, a couple assumptions must be kept:
  - groupings must be dichotomous (binning continous or ordinal variables)
  - groupings cannot change with time
- Note that the inability to use continuous predictors may not be ideal in instances where your predictor does not have a clear cut point. Usually you should try and find some sort of bimodal distribution, note that the groupings do not have to be the same size.

Cox proportional hazards model overview

- The model is expressed by the hazard function, $h(t)$:

$$h(t) = h_0(t) * \exp(b_1 x_1 + b_2 x_2 + b_3 x_3)$$

- Here, time is represented as t, and the b_n values are the coefficients for the covariates. $h_0(t)$ is the baseline hazard
- The function tells us the risk of the event at time t

Estimating a Cox Proportional Hazards model

- To estimate this I will use the `transplant` dataset in the `survival` package.
- My covariates of interest will be sex, age, and blood type.

```
data('transplant')
```

Estimating a Cox Proportional Hazards model

```
transplant<-subset(transplant,complete.cases(transplant))
transplant<-transplant %>% mutate(abo=ifelse(
    abo=="AB", "rare",'common'))%>%
  mutate(age=ifelse(age>=median(transplant$age),
                    "old","young"))%>%
  mutate(event=ifelse(event=="ltx"|event=="censored"|
                        event=="withdraw",F,T))
transplant$abo<-as.factor(transplant$abo)
transplant$age<-as.factor(transplant$age)
```

Estimating a Cox Proportional Hazards model

```
survivalobject<-Surv(time=transplant$futime,
                     event=transplant$event)
coxphmodel<-coxph(survivalobject ~ age + sex + abo,
                  data = transplant)
```

Estimating a Cox Proportional Hazards model

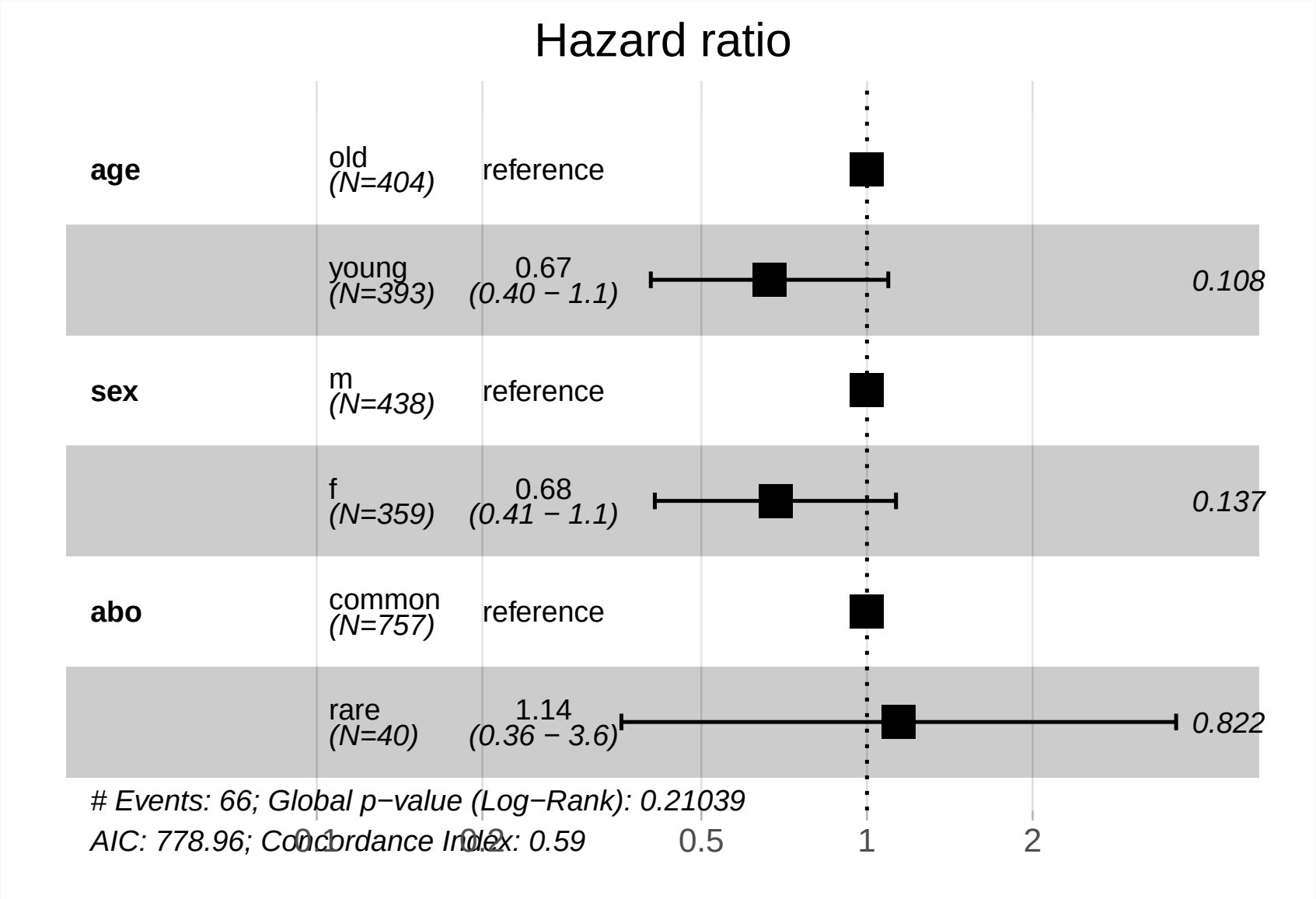
```
coxphmodel

## Call:
## coxph(formula = survivalobject ~ age + sex + abo, data = tran
##
##              coef exp(coef) se(coef)     z    p
## ageyoung -0.408     0.665    0.254 -1.61 0.11
## sexf     -0.383     0.682    0.257 -1.49 0.14
## aborare   0.133     1.142    0.592  0.22 0.82
##
## Likelihood ratio test=4.52  on 3 df, p=0.21
## n= 797, number of events= 66
```

- When reading CPH models: `exp(coef)` < 1 is a reduction of the hazard, `exp(coef)` > 1 is an increase in the hazard.

Estimating a Cox Proportional Hazards model

```
ggforest(coxphmodel, data = transplant)
```



Estimating a Cox Proportional Hazards model

- Similarly to the Kaplan-Meier curve for the Log-rank test, the visualization techniques make interpretation much easier.
- We can see that though none of these are significant, being younger and female are on the side of the vertical line that would indicate a reduction of hazard. Blood type is very close to 1 (no effect) and the intervals are impacted by the small number of cases with AB blood type.

Other types of Survival models

Overview of available models

- While the models discussed in this presentation are among the most commonly used, several others exist. Among them are the parametric models. If you suspect you data is distributed a certain way, these may be good to use. Common parametric models include:
    - Weibull
    - exponential
    - gamma
    - log-normal

Conclusion

In Conclusion

- Survival models are useful tools when your analysis focuses on some sort of event failure.
- While these models are most significantly used in the medical field, inferences based on politics can still be made.
- There are countless extensions to these models, so if you can think of a data type the follows the general trend, but do not think you could use one of these models presented today, there is probably one out there for you.