442-HW3

Yukun Gao

(I met professor at Friday's lecture and I requested for an extension until Nov. 23)

Summary

We want to find out whether individuals who use the left hand as the dominant hand are more susceptible to accidental death than people who use the right hand as the dominant hand. According to the results of Bayesian survival hierarchical model, we consider the hypothesis to be true. People who prefer to use the left hand tend to have less survival probability in accidental death than people who use the right hand. In addition, people who were born a century later tend to have a longer life. The reason could be: more advanced medical system and more friendly infrastructure designed for left-handed people improve longevity.

Introduction

For this assignment, we are particularly interested in whether the preference of the dominant hand has an impact on death due to accidents. To be more specific, whether people who use their left hand are more likely to die due to accidents compared with people who use the right hand. The data we use is called "cricketer data" which contains 5960 observations within DAAG package. The data contains information for each person's status(death or alive) as well as the three causes of death— kill in bad(natural death), accident and kill in action. In addition, we treat "kill in action" as an accident and treat people who died due to "kill in bed" as censoring.

Method and Model

Survival analysis is a branch of statistics for analyzing the expected duration of time until one event occurs. In our analysis, our particular interest is the lifetime of cricketers and thus, we could treat our data as survival data and we could apply survival analysis. We use "lifetime" to measure the time at which the accident death occurred and "lifetime=0" represents the time which a person was born. We treat "kill in bed" as censoring, and the person who is alive(not dead) as censoring, too. In this case, we should use Weibull distribution since while having only two parameters, it can properly describe the duration of time until death. Thus, we are going to fit the data with a Bayesian survival hierarchical model with censoring:

$$Z_i|Y_i, A_i = min(Y_i, A_i)$$
$$E_i|Y_i, A_i = I(Y_i < A_i)$$

"Kill in bed" is treated as censoring, which means we do not consider the person as "dead" when they actually die due to "kill in bed". Z_i represents the expected lifetime of ith cricketer. A_i represents the age of ith cricketer in 1992. Y_i is the lifetime of ith cricketer. E_i is the event indicator. If $E_i = 0$, this means the person died after the data is collected. If $E_i = 1$, this means the person died before the data is collected. Thus, if a person is "kill in bed", E_i should be 0 by our assumption.

In addition:

$$Y_i \sim Weibull(\lambda_i, \alpha)$$
$$\lambda_i = exp(-X_i\beta)$$

Y_i is the response variable which represents the lifetime of individual i. If $E_i = 0$, then the observed Y_i is not the real Y_i. We just assume that Y_i follows Weibull distribution with shape parameter α and scale parameter λ_i. X_i is the predictor and β is the corresponding parameters.

Furthermore, we should set a reasonable log-normal prior. Based on our data, the average life time is about 70 years old and by our experience, people are more likely to die around age 70. Therefore, it is reasonable to take log(7.5) as the mean of the log-normal distribution and $\frac{2}{3}$ as the standard deviation:

$$\alpha \sim LogNormal(\ log(7.5), \frac{2}{3})$$

Results:

Table 1 and Table 2 shows the estimated parameter given by our model as well as 95% CI. The number in Table 1 is shown in the exponential scale (exponentiated for the estimates) whereas Table 2 is in the percentage exponential scale. According to Table 1, the relative expected survival time for left-hand dominant individual $\exp(\beta_{\mathrm{i}})$ is about 0.724. This means that left-handed cricketers have 27.6% less lifetime on average than right-handed cricketers who died from accidents. This corresponds to the results shown in Table 2 for "left handed". The 95% CI is: (7.007% to 43.24%) which excludes 0(the positive number shows decrease in life time for left-handed people, not increase, don't get me wrong). This shows strong evidence that people w ho use left hand have a shorter lifespan than people who use right hand due to accidents. In addition, the estimate in percentage with exponential scale of "birth year" in Table 2 shows that people who born a century later increase lifespan by 75% more and the 95% CI is (23.09%, 153.17%,). The 95% CI excludes 0 which give us evidence that the people who born a century later tend to have a longer lifespan.

Table 1: Estimated parameter in exponential scale

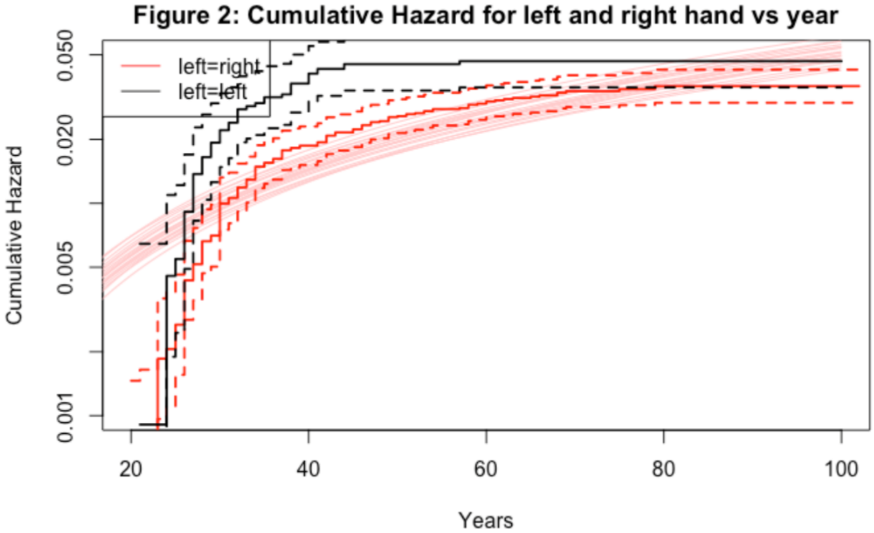
	Est	Lower	Upper
Shape	1.373	1.482	1.222
Reference(born 1900, right)	8.942	6.85	12.073
Birth year (per century)	1.75	1.231	2.532
Left handed	0.724	0.568	0.93

Table 2: Estimated parameter in percentage with exponential scale

	Est	Lower	Upper
Birth year(per century)	75.03772	23.08867	153.1655
Left handed	-27.58854	-43.23902	-7.006994

Figure 2 shows the cumulative hazard. Hazard could be interpreted as the rate of instantaneous events at a particular point in time t. In our case, hazard is accidental deaths. According to Figure 2, the black line represents the lifespan for people who use the left hand and the blue line represents the lifespan for people who use the right hand. The dashed line shows the confidence interval of the estimated cumulative hazard). The red lines in the middle show the sample cumulative hazard. We can see that for most of the time, the black lines stay above the red line. The confidence interval does not overlap a lot for "left hand" and "right hand". This confirms that people who use their left hand as the dominant hand are more likely to die from accidents compared to people who use the right hand.

Figure 2: Cumulative Hazard for left and right hand vs year



Limitation

There are many covariates, so purely making comparisons between left handers and right handers seems inadequate. Moreover, there are other factors that we don't consider which could be significant in affecting our result. The dataset itself might not be perfect and there might be some bias for the dataset.

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
#install.packages("dplyr")
library(dplyr)
#install.packages("survival")
library(survival)
#install.packages("survminer")
library(survminer)
#install.packages("INLA",repos=c(getOption("repos"),INLA="https://inla.r-inla-download.org/R/stable"),
library("INLA", quietly = TRUE)

#install.packages("DAAG")
library(DAAG)
data("cricketer", package = "DAAG")

cricketer$deadNotKia =
  as.numeric((cricketer$dead == 1) & (cricketer$inbed == 0))

cricketer$lifeC = cricketer$life / 100
cricketer$timeC = (cricketer$year - 1900)/100

cFitC = inla(inla.surv(lifeC, deadNotKia) ~ timeC + left, data=cricketer,
             family='weibullsurv',
             control.family = list(variant=1,
                                   hyper=list(alpha = list(prior = 'normal',
                                                           param = c(log(7.5),
                                                                     (2/3)^(-2)) ))),
             control.compute = list(config=TRUE),
             control.inla = list(strategy='laplace',
            fast=FALSE, h=0.0001), control.mode = list(theta = log(6),
                                                        restart=TRUE),verbose = TRUE)
knitr::kable(rbind(cFitC$summary.fixed[, c(1, 3, 5)],
                   cFitC$summary.hyper[, c(1, 3, 5)]), digits = 3)
xSeq = seq(0, 100, len = 200)
densHaz = Pmisc::sampleDensHaz(fit = cFitC, x = xSeq, n = 20, scale = 100)

hist(cricketer$life, xlab="lifetime", main="")
title(main="Figure 1: Histogram of lifetime")

hazEst = survfit(Surv(life, deadNotKia) ~ left, data = cricketer)
matplot(xSeq, densHaz[, "cumhaz", ],
        type = "l", lty = 1, col = "#FFCCCC", log = 'y',
        ylim = c(0.001, 0.05), xlim = c(20, 100),
xlab = "Years", ylab = "Cumulative Hazard",
main = "Figure 2: Cumulative Hazard for left and right hand vs year")
legend("topleft", col=c("red", "black"), legend=names(hazEst$strata), lty=1)
legend=gsub(',*=','',names(hazEst$strata))
lines(hazEst, fun = "cumhaz", col=c('red', 'black'), conf.int = TRUE, lwd = 1.8)


resTable = rbind(exp(-cFitC$summary.fixed[, c(4, 5, 3)]),
                 cFitC$summary.hyper[, c(4, 5, 3)])
```

```
rownames(resTable) = c('reference (born 1900, right)',
                       'birth year (per century)', 'left handed', 'shape')
colnames(resTable) = c('est','lower','upper')
knitr::kable(resTable[c(4,1,2,3),],digits=3)
100 * (resTable["birth year (per century)" ,] - 1)
100 * (resTable["left handed" ,] - 1)
```