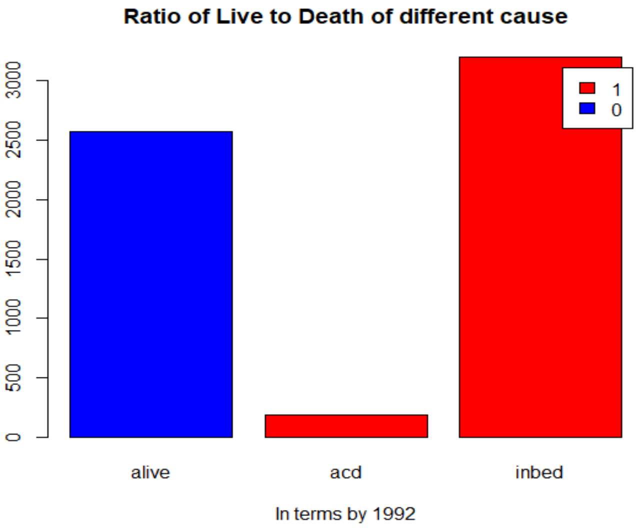
STA442 - Assignment 3 – Survival Analytics on Death Rate Due to Accident for British Cricketers

Jeff Xu
1003342545

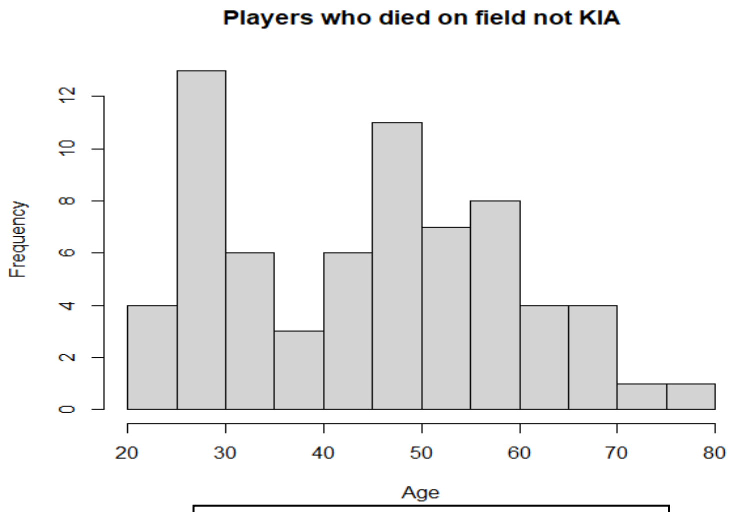
Introduction

in countries such as the UK where firearm is legal with a certificate, as firearms are made for righthanded users as the majority and clears bullet shells on the right side, which position where lefthanded user positions their face at, which poses danger to lefthanded users when using firearm in sports such as hunting and has a high chance of causing accidents. And although may not be as extreme, cases like this exists in many other sports that may leads to a higher accident rate for lefthanded individuals and a rise in their death rate due to their handedness.

Therefore, here we wish to raise the hypothesis that left-handed people are more susceptible to accidental death than right-handed people, and we will examine it through the data of a sport that is more mild - cricketing, with the data provided by DAAG Cricket data, which includes the year, birth, and lifespan, etc. of British first-class cricketers born in the time span of 1840-1960, and their leading causes of deaths of in bed due to sickness or other reasons, and the deaths by accidents which we wish to examine, with killed in action as a subset of those who died by accident as the time span covers events such as both world wars which we would wish to remove as we do not consider it as accident here, together with the information of their handedness included, as of 1992 with 5960 observations.



Bar plot for ratio of live to accidence deaths to inbed deaths



Histogram of frequency on players die on field not KIA

From above plot showing the ratio of life to accidence deaths to in bed deaths, we can tell that as of 1992, they are some cricketers that are still alive by the year 1992, and accidence together with killed in action as a cause of death for the cricketers are low comparatively to those who died in bed. And on the histogram of frequency on players die on field not KIA, it shows that most cricket players died of accidents in their late 20's, and with low death frequencies on their early 20's and 30~40's and have a trend of going high again after they passed the age of 40.

Model and Method

As we wish do conduct analytics on the comparison of the rate of accidental death with relation to their handedness, we would use a method and model as to conduct a survival analysis.

Which the lifetime as our outcome will be presented by Y_i so we choose to use a Weibull distribution as rate of survivals are usually left skewed as similar to Weibull, and we choose a INLA variant = 1 as it would only make a difference in terms of interpretation:

$$Y_i \sim Weibull(\lambda_i, \alpha)$$

Which is defined as follows:

$$\pi(x; \lambda_i, \kappa) = \frac{\kappa}{\lambda}(\frac{x}{\lambda})^{\kappa-1} \exp\,[-\left(\frac{x}{\lambda}\right)^{\kappa}]$$

Here with κ being the shape parameter, and λ_i is the scale parameter,

$$\lambda_i = \exp\,(-\eta_i)$$

$$\lambda_i = e^{-x\beta}$$

$$\eta_i = X_i\beta$$

And here we introduce the Censoring, which the sample is censored in that only the individuals were followed up by the study and survived up to the time frame covered. Which in this analytic, we have both of our events of KIA and Inbed being censoring as we only wish to consider the deaths caused by accidents, and of our samples who not dead by the end of time frame covered also censoring.

There we have E_i presenting the censor event which leads to the sample to die, with Y_i as the current shown lifetime of the sample, and their actual lifetime would be of their final age recorded in the data as Z_i, and A_i here presents the age of the individual cricketer. Which at $E_i = 1$, the sample is dead by an event, and $Y_i = Z_i$, and as $E_i = 0$, the sample is still alive and $Z_i < Y_i < \infty$, since his lifetime would be between his birthday assuming they are newborn, to their current age as they are still alive, to infinity as we don't know how long they are going to continue living as we don't know their survival after.

And the Censoring is as shown below as a hierarchical model:

$$Z_i | Y_i, A_i = \min\,(Y_i, A_i)$$

$$E_i | Y_i, A_i = \mathrm{I}\,(Y_i < A_i)$$

$$Y_i \sim Weibull[\,g(\eta_i, \kappa), \kappa\,]$$

$$\eta_i = X_i\beta$$

And the Link Function being:

$$g = e^{-X\beta} = e^{-\eta}$$

And as for the prior, $log{\sim}normal(\log(7.5,\frac{2}{3}))$ prior seems reasonable to be used in this analytic,

$$E(Y_i) = \lambda_i\Gamma(\frac{1+1}{\kappa})$$

$$\log{(\lambda_i)} = X_i\beta$$

With *i* presenting the left handedness, and *j* presenting the right handedness, then we can suppose that $X_{iq} = X_{jq}$ except that $X_{ip} = X_{jp} + 1$, and:

$$\frac{E(Y_i)}{E(Y_j)} = \frac{\exp{(X_i\beta)}\Gamma(1+1)/\kappa}{\exp{(\beta_p)}}$$

$$= \exp{(\beta_p)}$$

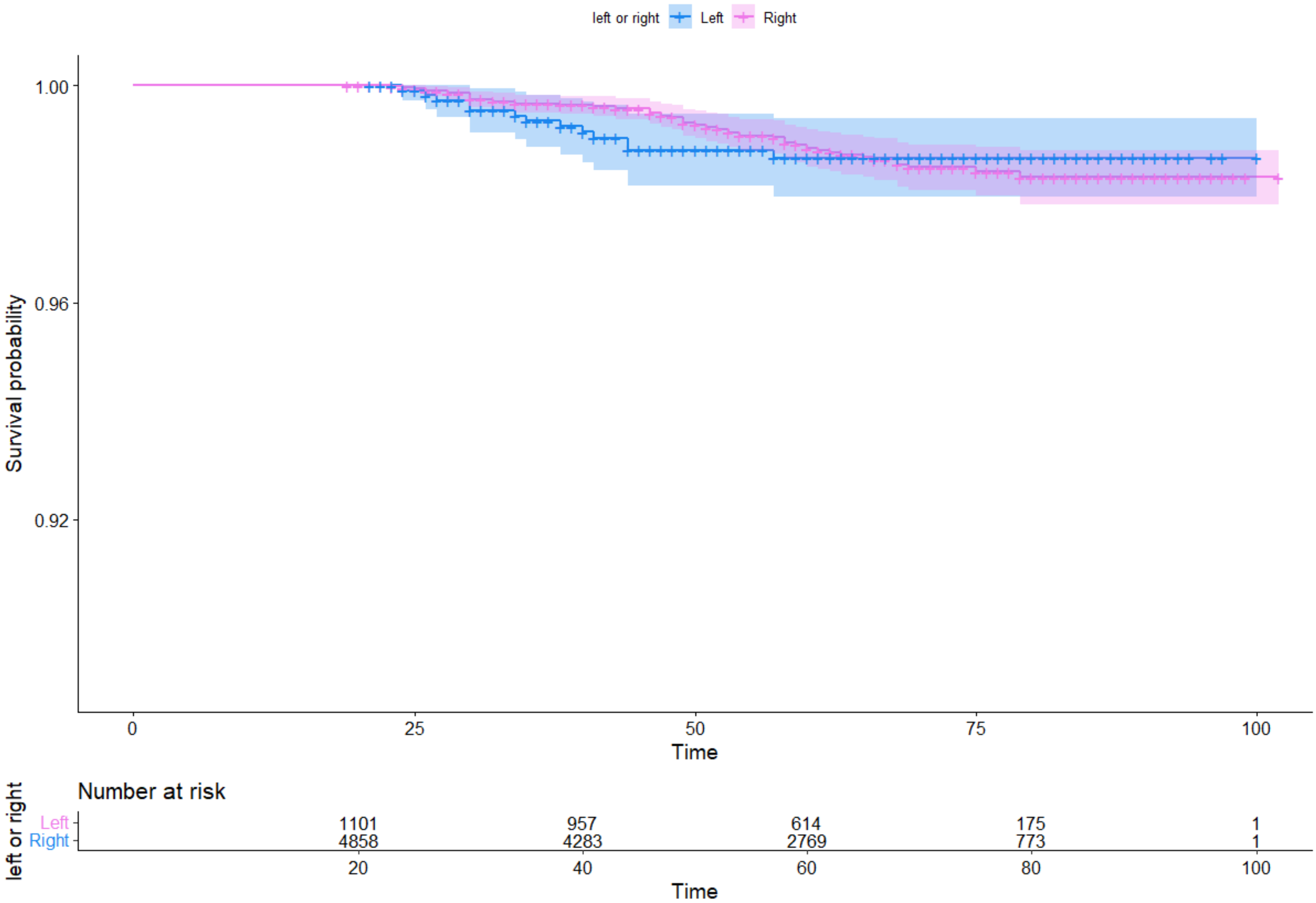
With the result $\exp{(\beta_p)}$ being the relative survival time expected, in a ratio of the lefthanded to righthanded.

Results

	Value	*Std. Error*	*z*	*p*
Intercept	6.268	0.232	27.066	0.000
leftleft	-0.028	0.145	0.846	0.846
Log(scale)	-0.756	0.109	-6.920	0.000

table for the summaries of fitted simple survival model

From above table for the summaries of fitted simple survival model, which as we did not use INLA here, and the results did not take the power of its exponential portion, it would be maybe to be not as accurate, so here we are only judging the coefficient of it to see if the impact of one's handedness is positive or negative. From the box highlighted red, we can tell being a lefthanded as a cricketer would have a negative impact on the lifetime / survival of the individual, and the rate tends to be fairly small. As we would know that A p-value less than 0.05 (typically ≤ 0.05) is statistically significant, and from the same row with the box highlighted red in above chart which shows the p-value is around 0.846, which implies that the impact on death by accident over being lefthanded is pretty insignificant.



Graph of survival probability for lefthanded and righthanded cricketers

The graph of survival probability above shows that, first, obviously, as am individual gets older with age getting larger, the survival probability becomes lower, and there is a difference between the survival chance of the lefthanded and righthanded, with the righthanded being of a higher rate of survival early on, but later falls off slightly after somewhere around 65~70 years old. The confident intervals are shown as lighter colored blocks, which we can see that the confident interval for the survival probability righthanded is more concentrated and thus precise while more disperse for the lefthanded. And for the numbers at risk are shown as some nice additional information for visualizing.

	est	*lower*	*upper*
birth year (change per century)	*1.299*	*0.738*	*2.325*
Left-handed's ratio to right	*0.969*	*0.636*	*1.558*
alpha	*1.364*	*1.531*	*1.212*

table of confidence interval for the relative expected survival rate for left handedness individuals of fitted INLA model

From above table shown with table shown above the relative expected survival rate for left handedness by the INLA model, which we should assume to be more accurate, we see that the rate of survival of the lefthanded to righthanded is 0.969:1, which gives us the result that the lefthanded people have 3.1% shorter lives than those who are righthanded. Which lies somehow in unison with our hypothesis, and the rate can be considered fairly minor and thus, insignificant as we have obtained from our analysis over the simple model. And We can also gain some other related information, such as the change of survival rater per the century, such that those who are born 100 years after those before them has a 29.9% increase in their life expectancy.

Summary

In this analytic we considered the fact that most of the products are produced with consideration for the righthanded majority people, we rise our hypothesis on that the minority lefthanded individuals might face a higher chance of accidents. We decide to use the cricket sport to conduct our analysis, with the data provided by DAAG cricketer data as of 1992 of first-class cricketers in the UK. First with findings that of the whole data set, there are still individual being sampled living by the end of the sampling period, and as for the causes of deaths accidents including being killed in action in the world wars or other conflicts are comparatively low as most pass away in bed. Thus, we decided to conduct our analysis with a method of survival analysis, with a simple one and one with INLA for better accuracy.

The result is as that firstly from the fitted simple model, it shows us that the lefthanded individuals does have a negative impact on their survival chance due to their handedness which lies in unison with our hypothesis but shown later by the p-value that the significance of this is fairly low and thus, it is an insignificant factor to the deaths of the cricketers. Then secondly from our fitted INLA for better accuracy, we also obtained that the righthanded individuals has a lower rate of lifetime / survival than the righthanded individuals, with a value about 3.1% lower, which matches the result from the simpler model and our hypothesis, and again it is fairly small and thus not as significant.

Appendix

Please refer to the R codes attached below:

```
################################################################################
############################# STA 442 ASSIGNMENT 3 ##############################
################################################################################

################################# LIBARIES ######################################

#install.packages("survminer",dependencies = TRUE, repos = 'http://cran.rstudio.com/')
####### WHATEVER packages needed
library(dplyr)
library(survival)
library(survminer)
library(INLA)
library(ggplot2)

############################### DATA ############################################
data("cricketer", package = "DAAG")
??cricketer
dat = cricketer
dat

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

################################################################################

#show why weibull distribution
#hist(dat$life, xlab = "lifetime",main = "")

#fit a model, no predictor
#dat$decade = (dat$year - 1850)/10
#dat$ones = 1
#cFitS = survreg(Surv(life, ones) ~ decade + left, data=dat, dist='weibull')
#knitr::kable(summary(cFitS)$table,digits=3)

#xSeq = seq(0,100,len=1000)
#plot(xSeq,dweibull(xSeq,shape = 1/cFitS$scale,scale =
exp(cFitS$coef['(Intercept)'])),xlab='age',ylab='dens', type='l')
#plot(xSeq,exp(cFitS$coef['(Intercept)']) * xSeq^(1/cFitS$scale) /cFitS$scale, type='l',
xlab='age',ylab='hazard')
```

```
################################ STATS INFOS ####################################


#deaths <- cricketer[which(cricketer$dead>0),]
#deaths <- cricketer[which(cricketer$dead>0),]
#qplot(life, data=deaths, xlab="Age")
#hist(cricketer$life, main= "Lifespan of all Players", xlab = "Age")

livedead = table(cricketer$dead, cricketer$cause)
barplot(livedead, main = "Ratio of Live to Death of different cause", xlab = "In terms by 1992",
    col=c('blue','red'), legend = rownames(livedead))

accidentonly <- cricketer[which((cricketer$inbed == 0)&(cricketer$kia == 0)&(cricketer$dead == 1)),]
hist(accidentonly$life, main= "Players who died on field not KIA", xlab = "Age")


############################## model with left ##################################
####### cricketer who are dead of accidents only

formula = survival::Surv(life, deadNotKia) ~ left
cFitS = survival::survreg(formula, data =cricketer,  dist = "weibull")



#summary
range(cricketer$life)
obj=Surv(time=cricketer$life, event=cricketer$deadNotKia )
sfit2 <- survfit(obj~cricketer$left)
summary(sfit2, times=seq(25, 100, 1))

# plots
plot(sfit2,ylim=c(0.89,1))

ggsurvplot(sfit2,data=dat,ylim=c(0.89,1))

ggsurvplot(sfit2, data=dat,ylim=c(0.89,1),
      conf.int=TRUE, pval=TRUE, risk.table=TRUE,
      legend.labs=c("Left", "Right"),
      legend.title="left or right",
      palette=c("dodgerblue2", "orchid2"),
      risk.table.height=.15,)

cricketer$lifeC = cricketer$life / 100
```

```
cricketer$timeC = (cricketer$year - 1900)/100

###tables
knitr::kable(summary(cFitS)$table, digits = 3)
knitr::kable(summary(sfit2)$table, digits = 3)



######################### FITTING MODEL WITH INLA #############################
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)

resTable = rbind(exp(-cFitC$summary.fixed[,c(4,5,3)]), cFitC$summary.hyper[, c(4, 5, 3)])
resTable #SHOW
rownames(resTable) = c('reference (born 1900, right)', 'birth year',"left","alpha")
colnames(resTable) = c('est','lower','upper')
knitr::kable(resTable[c(1,2,3,4),],digits=3)
100*(resTable["left", ] - 1)

############################### END OF CODES USED #############################

###GRAPH FOR HAZ
xSeq = seq(0, 100, len = 200)
densHaz = Pmisc::sampleDensHaz(fit = cFitC, x = xSeq, n=20, scale=100)
hazEst = survfit(Surv(life, event) ~ 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),lwd = 1, xlab = "Years", ylab = "Cumulative Hazard",
       main = "Graph 1: Cumulative Hazard Curve")
lines(hazEst, fun = "cumhaz", col=c('#660000', 'blue4'), lwd = 1.8)
legend("topleft", legend=c("Right-handed", "Left-handed"), col=c('#660000', 'blue4'), lty=1, cex=1,
      box.lty=0)
```