Análisis de sobrevivencia aplicado a base de datos de pacientes de sexo masculino diagnosticados con cáncer de laringe

Introducción

El modelo de Cox es un modelo muy útil para cuando se tienen datos de sobrevivencia de un estudio clínico; el objetivo de nosotros entonces es presentar este importante modelo para el análisis de datos de sobrevivencia, tomaremos dos ejercicios propuestos y vamos a realizar las estimativas de los parámetros y sus respectivos coeficientes, también haremos la adecuación del mejor modelo ajustado para el respectivo conjunto de datos en estudio. Veamos entonces el 1er ejercicio propuesto.

1) Para este ejercicio vamos a tomar los tiempos de 90 pacientes de sexo masculino diagnosticados en el periodo de 1970 a 1978 con cáncer de laringe que fueron acompañados hasta 01/01/1983. Para cada paciente fueron registrados, un diagnóstico, la edad (en años) y la etapa de la enfermedad (I= tumor primário, II= envolvimiento de nódulos, III=metástasis y IV= combinaciones de las 3 etapas anteriores), también sus respectivos tiempos de muerte o censura (en meses). Estas etapas se encuentran ordenadas por el grado de seriedad de la enfermedad (menos serio hasta más serio).

Utilizemos entonces el modelo de Cox para realizar el respectivo análisis sobre estos datos.

```
library("survival")
library("KMsurv")
library("survMisc")
library("survminer")
library("ggfortify")
library("flexsurv")
library("actuar")
library("dplyr")
library("asaur")
library("ranger")

#-------------------------------------------------------------------------------------------------
#-------------------------------------------------------------------------------------------------
#id =identificación del paciente; tempos = tiempo hasta la muerte (meses);
#cens = indicadora de censura (1 = falla e 0 = censura); estagio = etapa de la enfermedad
laringe<-read.table("laringe.txt",header=T) ###Datos utilizados en el estudio sobre el cáncer de laring
attach(laringe)
#-------------------------------------------------------------------------------------------------
#-------------------------------------------------------------------------------------------------

fit1<-coxph(Surv(tempos,cens)~factor(estagio), data=laringe,x= F, method="breslow")
summary(fit1)
fit1$loglik

fit2<- coxph(Surv(tempos,cens)~factor(estagio)+ idade, data=laringe, x = T, method="breslow")
summary(fit2)
fit2$loglik

fit4<-coxph(Surv(tempos,cens)~factor(estagio)+idade+factor(estagio)*idade, data=laringe, x = T, method=
```

```
summary(fit4)
fit4$loglik

TRV=2*(fit4$loglik[2]-fit2$loglik[2])
print(TRV)
gl=3
alpha=0.05
pvalue=pchisq(q=TRV,df=gl,ncp=0,lower.tail=FALSE,log.p=FALSE) #Valor p
pvalue
pvalue<alpha # no rechazo Ho
```

- Analizando el test de la razón de verosimilitud parcial, en el cual queremos saber si la interacción entre las variables edad y etapa es significativa, da como resultado TRV=6.20 y un valor p=0.10, con gl=3; dado que para un nivel de significancia $\alpha = 0.05$, $p > \alpha = 0.05$, esto indica que la interacción no es significativa. Por otro lado,el ajuste cuatro indica de que por lo menos uno de los parámetros betas asociados a la interacción difiere significativamente de cero, especificamente β_5 con valor p=0.022.

*Entonces, de los resultados encontrados, se decidio por la realización del análisis del modelo de Cox con interacción y sin interacción utilizando los residuos estandarizados de schoenfeld para luego proceder a escoger uno de los dos modelos.

```
#Evaluación del modelo de Cox ajustado sin interacción
residuos<-resid(fit2,type="scaledsch") #residuos de Schoenfeld estándarizados
cox.zph(fit2,transform="identity") #Tabla de los coeficientes de correlación entre los residuos de
# Schoenfeld estándarizados y la función  identidad g(t)=t, valor de la chi cuadrado
#y el valor-p

fit3<- coxph(Surv(tempos,cens)~factor(estagio)+ idade, data=laringe, x = T, method="breslow")
summary(fit3)
temp<-cox.zph(fit3, terms=FALSE)
print(temp)                    # display the results
par(mfrow=c(3,4))
plot(temp)
```

*Análisis: De los resultados se puede observar que el valor p de la covariable factor(estagio)3 es mayor que 0.05, pero muy cercano $\alpha = 0.05$.Esto indica que hay evidencias para decir que esta covariable no satisface la hipotesis de riesgos proporcionales.Los p valores para las demás covariables son mayores que 0.05, esto indica que no hay evidencias para rechazar la hipotesis de riesgos proporcionales para estas dos covariables. Por otro lado el valor p del test global para analizar la hipótesis general de riesgos proporcionales para todas las covariables conjuntamente en el modelo es mayor que alpha=0.05.Lo cual indica que no hay evidencias para rechazar la hipotesis de riesgos proporcionales para todas las covariables como un todo.

Así las cosas el modelo a considerar va a ser el modelo con interacción entre las variables edad y etapa, ya que con este modelo todas las covariables poseen riesgos proporcionales.

```
#Modelo final considerado
fit4<-coxph(Surv(tempos,cens)~factor(estagio)+idade+factor(estagio)*idade, data=laringe, x=T, method="b
summary(fit4)

#Estimativas de las funciones no paramétricas S0, h0, y Gamma0
#-------------------------------------------------------------------------------------------------
#-------------------------------------------------------------------------------------------------
ss<-survfit(fit4)
round(ss$surv,digits=5) # S(t|x) para x = xbar (default R) #
b<-fit4$coefficients
b<-as.vector(b)
```

```
b
x<- fit4$x
xbar<-as.matrix(apply(x,2,mean))
embx<-exp(-sum(b*xbar))
S0<-(ss$surv)^embx
H0<- -log(S0)
x1<-as.matrix(H0)
n<-nrow(x1)
a0<-rep(0,n)
for(i in 1:n){a0[i]<-H0[i+1]-H0[i]}
alpha0<-c(H0[1],a0[1:(n-1)])
alpha0<-c(H0[1],a0[1:(n-1)])
round(cbind(ss$time,S0,alpha0,H0),digits=5)
```

*Analizemos entonces las gráficas de las curvas de sobrevivencia estimadas para los pacientes con edades de 50 y 65 años para cada uno de las cuatro etapas de la enfermedad.

```
#Etapa I
S50I=(S0)^(exp(as.numeric(b[4])*50))
S65I=(S0)^(exp(as.numeric(b[4])*65))

#Etapa II

S50II=(S0)^(exp(as.numeric(b[1])+(as.numeric(b[4])+as.numeric(b[5]))*50))
S65II=(S0)^(exp(as.numeric(b[1])+(as.numeric(b[4])+as.numeric(b[5]))*65))

#Etapa III

S50III=(S0)^(exp(as.numeric(b[2])+(as.numeric(b[4])+as.numeric(b[6]))*50))
S65III=(S0)^(exp(as.numeric(b[2])+(as.numeric(b[4])+as.numeric(b[6]))*65))

#Etapa IV

S50IV=(S0)^(exp(as.numeric(b[3])+(as.numeric(b[4])+as.numeric(b[7]))*50))
S65IV=(S0)^(exp(as.numeric(b[3])+(as.numeric(b[4])+as.numeric(b[7]))*65))

par(mfrow=c(1,2))

#Gráfico edad 50 años para las cuatro etapas
plot(ss$time,S50I, lty=2, col=2, lwd=1.3)
lines(ss$time,S50II,lty=3, col=3,lwd=1.3)
lines(ss$time,S50III,lty=4, col=4,lwd=1.3)
lines(ss$time,S50IV, lty=5, col=5,lwd=1.3)
legend(7.0,0.8,lty=c(2,3,4,5), c("I", "II", "III", "IV"),col=c(2,3,4,5), cex=0.8, bty="n")

#Gráfico edad 65 años para las cuatro etapas
plot(ss$time,S65I, lty=2, col=2,lwd=1.3)
lines(ss$time,S65II,lty=3, col=3,lwd=1.3)
lines(ss$time,S65III,lty=4, col=4,lwd=1.3)
lines(ss$time,S65IV, lty=5, col=5, lwd=1.3)
legend(7.0,0.8,lty=c(2,3,4,5), c("I", "II", "III", "IV"),col=c(2,3,4,5), cex=0.8, bty="n")
```

*Análisis:de los dos gráficos de las curvas de sobrevivencia estimadas para los pacientes con edades de 50 y 65 años en cada uno de los cuatro estagios (etapa) de la enfermedad podemos observar que las curvas de sobrevivencia para los estagios I, III IV de los grupos con edades 50 años y 65 años no presentan diferencias

significativas en el estagio (etapa) dos se observan diferencias significativas en la función de sobrevivencia estimada para los grupos con edades de 50 y 65 años, para la curva de sobrevivencia del grupo de 65 años se observa un decrecimiento significativo en relacion a la curva de sobrevivencia del grupo con edad de 50 años. Esta observación gráfica justifica la presencia de la variable interacción principalmente la interacción entre Estagio II y Edad, ya que dependiendo de la edad.El efecto del estagio dos es diferente.

Ahora observemos las gráficas de las curvas de riesgo acumulado estimadas para los pacientes con edades de 50 y 65 años para cada uno de las cuatro etapas de la enfermedad, y saquemos nuestro respectivo análisis.

```
#Etapa I

H50I=(-log(S0))*(exp(as.numeric(b[4])*50))
H65I=(-log(S0))*(exp(as.numeric(b[4])*65))

#Etapa II

H50II=(-log(S0))*(exp(as.numeric(b[1])+(as.numeric(b[4])+as.numeric(b[5]))*50))
H65II=(-log(S0))*(exp(as.numeric(b[1])+(as.numeric(b[4])+as.numeric(b[5]))*65))

#Etapa III

H50III=(-log(S0))*(exp(as.numeric(b[2])+(as.numeric(b[4])+as.numeric(b[6]))*50))
H65III=(-log(S0))*(exp(as.numeric(b[2])+(as.numeric(b[4])+as.numeric(b[6]))*65))

#Etapa IV

H50IV=(-log(S0))*(exp(as.numeric(b[3])+(as.numeric(b[4])+as.numeric(b[7]))*50))
H65IV=(-log(S0))*(exp(as.numeric(b[3])+(as.numeric(b[4])+as.numeric(b[7]))*65))

#-------------------------------------------------------------------------------------------------------
#-------------------------------------------------------------------------------------------------------
par(mfrow=c(1,2))
#Gráfico edad 50 años para las cuatro etapas
plot(ss$time,H50I, lty=2, col=2, lwd=1.3, ylim=c(0,4.5))
lines(ss$time,H50II,lty=3, col=3,lwd=1.3)
lines(ss$time,H50III,lty=4, col=4,lwd=1.3)
lines(ss$time,H50IV, lty=5, col=5,lwd=1.3)
legend(1.0,4.0,lty=c(2,3,4,5), c("I", "II", "III", "IV"),col=c(2,3,4,5), cex=0.8, bty="n")
#-------------------------------------------------------------------------------------------------------
#-------------------------------------------------------------------------------------------------------
#Gráfico edad 65 años para las cuatro etapas
plot(ss$time,H65I, lty=2, col=2,lwd=1.3, ylim=c(0,4.5))
lines(ss$time,H65II,lty=3, col=3,lwd=1.3)
lines(ss$time,H65III,lty=4, col=4,lwd=1.3)
lines(ss$time,H65IV, lty=5, col=5, lwd=1.3)
legend(1.0,4.0,lty=c(2,3,4,5), c("I", "II", "III", "IV"),col=c(2,3,4,5), cex=0.8, bty="n")
```

*Notemos que de las gráficas de tasa de riesgo acumulado estimadas, la tasa de riesgo acumulado en la etapa 4 para ambos grupos es muy alta en comparación con las otras tasas de riesgo acumulado, pero para el grupo de 65 años la tasa de riesgo acumulado en la etapa 4 es mucho mayor que la tasa de riesgo en el grupo de 50 años, también notemos que la tasa de riesgo acumulado para el grupo de 50 años en la etapa 2 es menor que la tasa de riesgo acumulado para el grupo de 65 años en la misma etapa.

Para terminar observemos las gráficas de las curvas de riesgo estimadas para los pacientes con edades de 50 y 65 años para cada uno de las cuatro etapas de la enfermedad, y saquemos nuestro respectivo análisis.

```
#Etapa I
h50I=sort(alpha0)*(exp(as.numeric(b[4])*50))
h65I=sort(alpha0)*(exp(as.numeric(b[4])*65))

#Etapa II

h50II=sort(alpha0)*(exp(as.numeric(b[1])+(as.numeric(b[4])+as.numeric(b[5]))*50))
h65II=sort(alpha0)*(exp(as.numeric(b[1])+(as.numeric(b[4])+as.numeric(b[5]))*65))

#Etapa III

h50III=sort(alpha0)*(exp(as.numeric(b[2])+(as.numeric(b[4])+as.numeric(b[6]))*50))
h65III=sort(alpha0)*(exp(as.numeric(b[2])+(as.numeric(b[4])+as.numeric(b[6]))*65))

#Etapa IV

h50IV=sort(alpha0)*(exp(as.numeric(b[3])+(as.numeric(b[4])+as.numeric(b[7]))*50))
h65IV=sort(alpha0)*(exp(as.numeric(b[3])+(as.numeric(b[4])+as.numeric(b[7]))*65))

#-------------------------------------------------------------------------------------------------------

par(mfrow=c(1,2))
#Gráfico edad 50 años para las cuatro etapas
plot(ss$time,h50I, lty=2, col=2, lwd=1.3)
lines(ss$time,h50II,lty=3, col=3,lwd=1.3)
lines(ss$time,h50III,lty=4, col=4,lwd=1.3)
lines(ss$time,h50IV, lty=5, col=5,lwd=1.3)
legend(1.3,0.08,lty=c(2,3,4,5), c("I", "II", "III", "IV"),col=c(2,3,4,5), cex=0.8, bty="n")

#Gráfico edad 65 años para las cuatro etapas
#-------------------------------------------------------------------------------------------------------
plot(ss$time,h65I, lty=2, col=2,lwd=1.3)
lines(ss$time,h65II,lty=3, col=3,lwd=1.3)
lines(ss$time,h65III,lty=4, col=4,lwd=1.3)
lines(ss$time,h65IV, lty=5, col=5, lwd=1.3)
legend(1.3,0.08,lty=c(2,3,4,5), c("I", "II", "III", "IV"),col=c(2,3,4,5), cex=0.8, bty="n")
```

- de las dos gráficas construidas para los grupos de edades de 50 y 65 años, notemos que la tasa de riesgo para el grupo de 50 años en la etapa 2 es menor que la tasa de riesgo para el grupo de 65 años en la misma etapa. Esto sigue dando clara evidencia de la interacción que hay entre la etapa 2 y la edad de los pacientes.

2) Para este ejercicio vamos a tomar la base de datos ashkenazi, la cual contiene datos de un estudio a judíos asquenazíes y cáncer de mama. El subconjunto consta de parejas de parientes femeninas de primer grado que también son parientes de primer grado de un probando. El formato en el que se presenta esta base de datos es Un marco de datos con 3920 observaciones sobre las siguientes 4 variables. famID(identificación familiar),brcancer(1 si el sujeto tenía cáncer de mama, 0 si no),age(Edad al inicio del cáncer de mama o edad actual si no hay cáncer de mama) y mutant(1 si el probando relativo de primer grado era portador de la mutación BRCA, 0 si no).

```
data("ashkenazi")
temp<-ashkenazi$age
cens<-ashkenazi$brcancer
attach(ashkenazi)
fit1_1<-coxph(Surv(temp,cens)~factor(ashkenazi$mutant), data=ashkenazi,x= F, method="breslow")
```

```
summary(fit1_1)
fit1_1$loglik

fit2_2<- coxph(Surv(temp,cens)~factor(ashkenazi$mutant)+ashkenazi$famID , data=ashkenazi, x = T, method=
summary(fit2_2)
fit2_2$loglik

fit4_4<-coxph(Surv(temp,cens)~factor(ashkenazi$mutant)+ashkenazi$famID+factor(ashkenazi$mutant)* ashkena
summary(fit4_4)
fit4_4$loglik

TRV=2*(fit4_4$loglik[2]-fit2_2$loglik[2])
print(TRV)
gl=1
alpha=0.05
pvalue=pchisq(q=TRV,df=gl,ncp=0,lower.tail=FALSE,log.p=FALSE) #Valor p
pvalue
pvalue<alpha # no rechazo Ho
```

*Analizando el test de la razón de verosimilitud parcial, en el cual queremos saber si la interacción entre las variables mutant y famId es significativa, da como resultado TRV=0.628 y un valor p=0.428, con gl=1; dado que para un nivel de significancia $\alpha = 0.05$, $p > \alpha = 0.05$, esto indica que la interacción no es significativa.

*Tomemos entonces el modelo sin interacción y veamos si este cumple con el supuesto de riesgos proporcionales utilizando los residuos estandarizados de Schoenfeld estándarizados y la función identidad g(t)=t,la cual nos dará el valor de la chi cuadrado y el valor-p, para cada uno de los parámetros asociados en la prueba

```
#Evaluación del modelo de Cox ajustado sin interacción
residuos<-resid(fit2_2,type="scaledsch") #residuos de Schoenfeld estándarizados
cox.zph(fit2_2,transform="identity") #Tabla de los coeficientes de correlación entre los residuos de
# Schoenfeld estándarizados y la función  identidad g(t)=t, valor de la chi cuadrado
#y el valor-p

fit3_3<- coxph(Surv(temp,cens)~factor(ashkenazi$mutant)+ ashkenazi$famID, data=ashkenazi, x = T, method=
summary(fit3_3)
temp1<-cox.zph(fit3_3, terms=FALSE)
print(temp1)                  # display the results
par(mfrow=c(3,4))
plot(temp1)
```

*Análisis: notemos que los p valores para todas las covariables son mayores que 0.05, esto indica que no hay evidencias para rechazar la hipotesis de riesgos proporcionales para estas dos covariables. Por otro lado el valor p del test global para analizar la hipótesis general de riesgos proporcionales para todas las covariables conjuntamente en el modelo es mayor que alpha=0.05.Lo cual indica que no hay evidencias para rechazar la hipotesis de riesgos proporcionales para todas las covariables como un todo.

Así las cosas el modelo a considerar va a ser el modelo sin interacción entre las variables mutant y famID, ya que con este modelo todas las covariables poseen riesgos proporcionales.

```
#Modelo final considerado
fit3_3<- coxph(Surv(temp,cens)~factor(ashkenazi$mutant)+ ashkenazi$famID, data=ashkenazi, x = T, method=
summary(fit3_3)
#Estimativas de las funciones no paramétricas S0, h0, y Gamma0
#-------------------------------------------------------------------------------------------------
#-------------------------------------------------------------------------------------------------
ss<-survfit(fit3_3)
```

```
round(ss$surv,digits=5) # S(t|x) para x = xbar (default R) #
b<-fit3_3$coefficients
b<-as.vector(b)
b
x<- fit3_3$x
x
xbar<-as.matrix(apply(x,2,mean))
embx<-exp(-sum(b*xbar))
S0<-(ss$surv)^embx
H0<- -log(S0)
x1<-as.matrix(H0)
n<-nrow(x1)
a0<-rep(0,n)
for(i in 1:n){a0[i]<-H0[i+1]-H0[i]}
alpha0<-c(H0[1],a0[1:(n-1)])
alpha0<-c(H0[1],a0[1:(n-1)])
round(cbind(ss$time,S0,alpha0,H0),digits=5)
```

*Dado que la variable mutant es significativa en el modelo cuando su valor es 1,entonces analizemos las gráficas de las curvas de sobrevivencia estimadas para los pacientes con identificador de familia de 87 y 223, con la variable mutant=1 .

```
#mutant=1

S87I=(S0)^(exp(as.numeric(b[1])+as.numeric(b[2]))*87)
S223I=(S0)^(exp(as.numeric(b[1])+as.numeric(b[2]))*223)


par(mfrow=c(1,2))

#Gráfico famID 87 para mutant=1
plot(ss$time,S87I, lty=2, col=2, lwd=1.3)
legend(7.0,0.8,lty=c(2,3,4,5), c("1"),col=c(2), cex=0.8, bty="n")

#Gráfico famID 223 para mutant=1
plot(ss$time,S223I, lty=2, col=2,lwd=1.3)
legend(7.0,0.8,lty=c(2,3,4,5), c("2"),col=c(3), cex=0.8, bty="n")
```

*Del gráfico observamos que la variable mutant para las familias identificadas con 87 y 223 no parece tener una diferencia muy significativa, para ambas familias parecen tener el mismo patrón de decaimiento, esto tal vez se deba a la falta de interacción entre la variable mutant con famID.

Ahora observemos las gráficas de las curvas de riesgo acumulado estimadas para los pacientes con identificación familiar de 87 y 223 y variable mutant=1

```
#mutant=1

H87I=(-log(S0))*(exp(as.numeric(b[1])+as.numeric(b[2]))*87)
H223I=(-log(S0))*(exp(as.numeric(b[1])+as.numeric(b[2]))*223)

#-------------------------------------------------------------------------------------------------------
#-------------------------------------------------------------------------------------------------------
par(mfrow=c(1,2))
#Gráfico famID 87 para mutant=1
plot(ss$time,H87I, lty=2, col=2, lwd=1.3, ylim=c(0,4.5))
```

```
legend(1.0,4.0,lty=c(2,3,4,5), c("I"),col=c(2), cex=0.8, bty="n")
#-------------------------------------------------------------------------------------------------------
#-------------------------------------------------------------------------------------------------------
#Gráfico famID 223 para mutant=1
plot(ss$time,H223I, lty=2, col=2,lwd=1.3, ylim=c(0,4.5))
legend(1.0,4.0,lty=c(2,3,4,5), c("I"),col=c(3), cex=0.8, bty="n")
```

*Notemos que de las gráficas de tasa de riesgo acumulado estimadas,son ambas muy parecidas si importar la identificación de la familia, ocurre otra vez un resultado muy parecido al que observamos en las curvas de sobrevivencia estimadas, ambas gráficas presentan en este caso una tendencia muy similar a medida que se van acumulando.

*Conclusión: la variable mutant cuando toma el valor de 1 (si el probando relativo de primer grado era portador de la mutación BRCA) es la más significativa en el estudio, esto tal vez se deba a la importancia que tiene el pariente principal el cual es portador de la mutación BRCA con el individuo en el estudio, pero la variable famId no tiene ninguna significancia en el modelo de Cox.