Final project

北大統計所碩二 郭翊萱

#---------------------------------------## Q1 ##---------------------------------------

# Researchers are interested in the age at which liver cancer is commonly diagnosed

getwd()

setwd("/Users/guoyixuan/Documents/surv\_final/")

data1 <- read.csv("/Users/guoyixuan/Documents/surv\_final/data\_HW4\_1.csv")

head(data1) # 2000 9

# ID, Age, follow\_up\_even\_time, censord\_index, HBSAG, GENDER, alcohol, alt1, alt2

# (時間變數值) follow\_up\_even\_time

# (事件變數值) censord\_index: whether cancer is present (0/1)

# HBSAG 是否是B肝帶原者: whether the individual is a carrier of hepatitis B (0/1)

# GENDER 性別: (0/1)

# alcohol 飲酒習慣: alcohol consumption (0/1)

# elevated mild liver index (0/1)

# elevated moderate to high liver index (0/1)

# (1) ------------------------------------------

# 使用Cox proportional hazard model中的partial likelihood部分似然來估計模型參數，可以解決如果資料中可能存在truncation的問題

# (2) ------------------------------------------

# 可以，年齡常常被當成影響存活曲線的指標

# (3) ------------------------------------------

library(survival)

fit = coxph(Surv(follow\_up\_Even\_time, Censord\_index) ~ Age+HBSAG+GENDER+alcohol1+alt1+alt2, data = data1)

fit

plot(survfit(fit))

# 是，從exp(coef) 的值可以看出，年齡對癌症具有統計顯著

# (4) ------------------------------------------

# harmful effect，每增加一年齡單位，會比前一個單位多1.1倍的風險得到癌症

# (5) ------------------------------------------

# 是，可以使用Cox model，他一般被用來推估共變數（年齡）對存活時間的影響，也可以用來預測特定時間的存活機會

#---------------------------------------## Q2 ##---------------------------------------

# 檢測incidence rate of cancer 跟 gender之間的關聯

data2 <- read.csv("/Users/guoyixuan/Documents/surv\_final/data\_HW4\_2.csv")

head(data2) # 2000 4

# (6)(7)(8)

fit2 = coxph(Surv(follow\_up\_Even\_time, Censord\_index) ~ GENDER, data = data2)

fit2

# 從此處可知，性別對是否得癌症是統計顯著的，由於男生是1女生是0，所以男生比起女生得到癌症的機率會大2.25倍

# (9)

# Generate survival curves

survival\_curves <- survfit(fit2)

plot(survfit(fit2))

# Extract 5-year survival rates for each gender

survival\_at\_5\_years <- summary(survival\_curves, times = 5)

survival\_at\_5\_years

#---------------------------------------## Q3 ##---------------------------------------

library(Copula.surv)

data3 = read.table("/Users/guoyixuan/Documents/surv\_final/data\_HW3.txt", header = TRUE)

head(data3) # 5000 4

# liver\_cancer

# dead\_time

# cencered\_index(1/0:no/yes)

# HBV(1/0:carrier/healthy)

# 研究者關心的是癌症發生的機率而不是死亡率

T1 = data3$liver\_cancer

T2 = data3$dead\_time

Z = data3$HBV

Z

typeS = names(table(data3$HBV))

Z[data3$HBV == typeS[1]] = 0

Z[data3$HBV == typeS[2]] = 1

# HBV = 0

x.obs\_0 = data3$liver\_cancer[Z==0]

y.obs\_0 = data3$dead\_time[Z==0]

dy\_0 = data3$cencered\_index[Z==0]

dx\_0 = ifelse(x.obs\_0 == y.obs\_0, 0, 1)

# HBV = 1

x.obs\_1 = data3$liver\_cancer[Z==1]

y.obs\_1 = data3$dead\_time[Z==1]

dy\_1 = data3$cencered\_index[Z==1]

dx\_1 = ifelse(x.obs\_1 == y.obs\_1, 0, 1)

#

z.obs = ifelse(data3$liver\_cancer <= data3$dead\_time, data3$liver\_cancer, data3$dead\_time)

dz = data3$cencered\_index

dz1 = which(data3$liver\_cancer < data3$dead\_time)

dz[dz1] = 1

z.obs\_0 = z.obs[Z==0]

z.obs\_1 = z.obs[Z==1]

dz\_0 = dz[Z==0]

dz\_1 = dz[Z==1]

theta\_0 = U2.Clayton(x.obs\_0, y.obs\_0, dx\_0, dy\_0) # theta\_0 = 42.158

theta\_1 = U2.Clayton(x.obs\_1, y.obs\_1, dx\_1, dy\_1) # theta\_1 = 15.09

Sz0 = survfit(Surv(z.obs\_0, dz\_0) ~ 0) #4762

Sz0

Sz1 = survfit(Surv(z.obs\_1, dz\_1) ~ 1, conf.type = "log-log") #238

Sz1

Sy0 = survfit(Surv(y.obs\_0, dy\_0) ~ 0, conf.type = "log-log") #4762

Sy0

Sy1 = survfit(Surv(y.obs\_1, dy\_1) ~ 1, conf.type = "log-log") #238

Sy1

pLA=function(yy, tt, LL){

pla=function(yi){

loc=sum(yi>=tt)

if(loc==0)ans=1

if(loc>0)ans=LL[loc]

ans

}

apply(matrix(yy),1,pla)

}

p.time = seq(1,30,by=0.5)

Sz0.y = pLA(p.time, summary(Sz0)$time, summary(Sz0)$surv)

Sz1.y = pLA(p.time, summary(Sz1)$time, summary(Sz1)$surv)

Sy0.y = pLA(p.time, summary(Sy0)$time, summary(Sy0)$surv)

Sy1.y = pLA(p.time, summary(Sy1)$time, summary(Sy1)$surv)

# survival function

Sx0.y = (Sz0.y^(-theta\_0[1])-Sy0.y^(-theta\_0[1])+1 )^(1/(-theta\_0[1]))

Sx1.y = (Sz1.y^(-theta\_1[1])-Sy1.y^(-theta\_1[1])+1 )^(1/(-theta\_1[1]))

# (10)

plot(Sx0.y, type = "s")

points(Sx1.y, type = "s", col = 2)

# HBV是0時(healthy)，存活的時間較長，罹癌的風險較低

# HBV是1時，存活曲線很快落至0，罹癌風險較高（存活機率較低）

# (12)

# 因為要分別計算HBV是0跟1對存活與否（是否罹癌）的影響，因此其cox model的模型參數會有所不同，

# 將其分層可以更好地捕捉到不同情況下的存活情況

# (13)

# the cox proportional hazards assume that the hazard ratio between any two groups is constant over time.

# (14)

# the pla function is used to approximate the survival function of T1 based on the survival function of other related time: T2<

# the pla function helps estimate the survival function of T1 for both HBV states, considering

# the limitations in the available data and copula model

---------------------------------------## Q4 ##---------------------------------------

rm(list=ls(all=TRUE))

data = read.table("/Users/guoyixuan/Documents/surv\_final/DATA\_4.txt"

, head = TRUE )

head(data)

library(survival)

# X1-發生肝纖維化(肝硬化)的時間(時間一般很長)

# X2:發生肝癌的時間(survival time一般只有5.6年)

# 一般先發生肝纖維化再發生肝癌(反過來很少見，而且肝癌survival time較短，常常等不到發生肝硬化)

# 通常有兩種情況：

# (1) 經過肝硬化再得到肝癌

# Health ------ popuulation\_T1 ------ population\_T2

# (2) 直接得到肝癌

# Health ---------------- population\_T2

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

# (1) 經過肝硬化再得到肝癌

# 競爭狀態：使用multistat model

p.time = seq(1, 30, by = 0.5)

T1 = data$X1 #肝纖維化的時間

T2 = data$X2 #肝硬化的時間

d2 = data$D # 現在這個人是不是有得到肝癌 # delta\_2

d1 = (T1<T2) # TRUE/FALSE # delta\_1 (X=T1)

# 其中若：

# X1=26.827737, X2=26.82774, D=0：還沒得到肝癌就離開實驗了(健康的人)(#9177)

# X1=18.8,X2=19.1,D=1：在X2得到肝癌，也得到肝硬化

# X1=19.1,X2=19.1,D=1：沒有肝硬化直接肝癌

Z = data$S

X = cbind(data$age, data$sex, data$smoking, data$alcohol, data$alt1)

summary(X)

m = length(T1)

m # 5000 資料中，censor rate 高達 0.9998 (很正常)

mean(d2) # 0.0112 # 肝癌的均數

mean(d1) # 0.0304 # 肝硬化的均數

# 約有1%得到肝癌，有3%得到肝硬化

# 研究目的：肝癌是很重要的國民疾病，今天搜集到一個資料，希望研究是否肝癌的發生率在給定有無肝硬化的情況下會不同，

# 而資料中有肝硬化的時間、C型肝炎的表面抗原

# 年齡(age)、性別(sex)、抽菸(smoking)(0:無)、飲酒習慣(alcohol)(0:無)、肝指數(alt1)(0:正常 1:輕度過標)

# 資料名稱：R資料

MM = as.matrix(cbind(Z, X))

colnames(MM) = c("C型肝炎的表面抗原", "age", "性別", "抽菸", "飲酒習慣", "肝指數")

MM

Lev\_M = apply(MM, 2, median)

BB1 = coxph(Surv(T1, d1)~MM)

#names(BB1$coef) = names(data)[4:9]

#names(BB1$coef)[1] = "HCV"

BB1

# 根據表格中exp(coef)的值，總共有3個顯著的因子

# 第一個顯著因子是C型肝原帶原者比起非帶原者得到肝硬化風險的指數比增高了7.3倍

# 第二個顯著因子是年齡，每增加一年齡單位，會比前一個單位多1.037倍的風險得到肝硬化

# 最後一個顯著的因子是肝發炎指數，可知每增加一個單位的肝指數發炎狀況，會比前一個單位多2.103倍的風險得到肝硬化

# 此外，性別、抽菸、喝酒從表格來看並無統計顯著

st1 = sort(T1, index.return = TRUE)

sd1 = d1[st1$ix]

MMst1 = MM[st1$ix, ] # 排序後的covariate

st1 = st1$x # 排序後的T1

exB1 = exp(apply(t(MMst1) \* BB1$coef, 2, sum))

exB1

dLA1 = sd1/(cumsum(exB1[m:1])[m:1]) # lambda\_1(t\_j)(見照片)

LA1 = cumsum(dLA1) # 累加所有的lambda\_d(t\_j)

# LA1\_i = pLA(T1, st1, LA1)

# plot survival curve

S1 = exp(-LA1)

S1\_HBV = exp(-LA1\*exp(BB1$coef[1]))

S1\_age = exp(-LA1\*exp(BB1$coef[2]\*48)) # coef \* median

S1\_alt = exp(-LA1\*exp(BB1$coef[6]))

plot(st1, S1, type = "s", ylim=c(0.96, 1)) # health

points(st1, S1\_HBV, type = "s", col = "2") #

points(st1, S1\_age, type = "s", col = "3")

points(st1, S1\_alt, type = "s", col = "4")

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

# (2) 直接得到肝癌

# 沒有經過肝硬化直接從健康到得到癌症 ((1-delta\_1)\*delta\_2)

# 只看罹癌者

T02 = ifelse(d1 == 1, T1, T2) # 把經過肝硬化的人宣告成censor

D02 = ifelse(d1 == 1, 0, d2) # 新的censor indicator

BB2 = coxph(Surv(T02, D02) ~ MM)

#names(BB2$coef) = names(data)[4:9]

#names(BB2$coef)[1] = "HCV"

BB2

# 顯著的剩下年齡跟C型肝原的表面抗原

# 對於是否得癌症，年齡是顯著的有害因子

# C型帶原者相較於沒帶原者罹癌風險增加5.04倍

# 每增加一歲，相較於前一歲罹癌的風險增加1.09倍

# 其他因子對罹癌風險的影響則無統計顯著

st2 = sort(T02, index.return = TRUE)

sd2 = D02[st2$ix]

MMst2 = MM[st2$ix, ]

st2 = st2$x

exB2 = exp(apply(t(MMst2) \* BB2$coef, 2, sum))

exB2

dLA2 = sd2/(cumsum(exB2[m:1])[m:1]) # lambda\_1(t\_j)(見照片)

LA2 = cumsum(dLA2) # 累加所有的lambda\_d(t\_j)

pLA = function(yy, tt, LL){

pla = function(yi){

loc = sum(yi >= tt)

if(loc==0)ans=0 # ans- hazard:0, survival:1

if(loc>0)ans=LL[loc]

ans

}

apply(matrix(yy), 1, pla)

}

LA2 = pLA(p.time, st2, LA2)

# plot survival curve

S2 = exp(-LA2)

S2\_HBV = exp(-LA2\*exp(BB2$coef[1]))

S2\_age = exp(-LA2\*exp(BB2$coef[2]\*48)) # coef \* median

plot(p.time, S2, type = "s", ylim=c(0.99, 1)) # health

points(p.time, S2\_HBV, type = "s", col = "2")

points(p.time, S2\_age, type = "s", col = "3")

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

# (3) 看肝硬化的部分

# left truncation的問題

LN1 = T2 > T1 # 一定抓到肝硬化的人

M3 = cbind(Z, X)[LN1, ]

D03 = (d1\*d2)[LN1]

T3 = T2[LN1]

R3 = T1[LN1] # 修正得到肝硬化的時間點

BB3 = coxph(Surv(R3, T3, D03) ~ M3)

BB3

# 從表格可以看出，只有C型肝原的表面抗原是統計顯著的，其他都非統計顯著

st3 = sort(T2, index.return = TRUE)

sd3 = (d1\*d2)[st3$ix]

sr3 = T1[st3$ix]

MMst3 = MM[st3$ix, ]

st3 = st3$x

m3 = length(st3)

exB3 = exp(apply(t(MMst3) \* BB3$coef, 2, sum))

R3m = matrix(sr3, m3, m3)

T3m = matrix(st3, m3, m3, byrow = TRUE)

Risk3 = ifelse((R3m <= T3m) & (T3m <= t(T3m)), 1, 0)

#dLA3 = sd3/(cumsum(exB3[m:1])[m:1]) # lambda\_1(t\_j)

dLA3 = sd3/apply(Risk3\*exB3, 2, sum)

LA3 = cumsum(dLA3) # 累加所有的lambda\_d(t\_j)

#LA3 = pLA(p.time, st3, LA3)

LA3\_i = pLA(p.time, st3, LA3)

LA3\_m = LA3\_i\*(exp(sum(BB3$coef\*Lev\_M)))

LA2\_m = LA2\*(exp(sum(BB2$coef\*Lev\_M)))

# plot

#plot(p.time, LA3\_m, type = "s")

#points(p.time, LA2\_m, type = "s", col = 2)

# plot survival curve

dLA3 = sd3/(cumsum(exB3[m:1])[m:1])

LA3 = cumsum(dLA3)

LA3 = pLA(p.time, st3, LA3)

S3 = exp(-LA3)

S3\_HBV = exp(-LA3\*exp(BB3$coef[1]))

S3\_age = exp(-LA3\*exp(BB3$coef[2]\*48)) # coef \* median

plot(p.time, S3, type = "s", ylim=c(0.99, 1)) # health

points(p.time, S3\_HBV, type = "s", col = "2")

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

LN1 = T2 > T1

M3 = cbind(Z, X)[LN1, ]

D03 = (d1\*d2)[LN1]

T3 = T2[LN1]

R3 = T1[LN1]

BB3 = coxph(Surv(R3, T3, D03) ~ M3)

st3 = sort(T2, index.return = TRUE)

sd3 = (d1\*d2)[st3$ix]

sr3 = T1[st3$ix]

MMst3 = MM[st3$ix, ]

st3 = st3$x

m3 = length(st3)

exB3 = exp(apply(t(MMst3) \* BB3$coef, 2, sum))

R3m = matrix(sr3, m3, m3)

T3m = matrix(st3, m3, m3, byrow = TRUE)

Risk3 = ifelse((R3m <= T3m) & (T3m <= t(T3m)), 1, 0)

dLA3 = sd3/apply(Risk3\*exB3, 2, sum)

LA3 = cumsum(dLA3)

LA3 = pLA(p.time, st3, LA3)

LA3\_m = LA3\*(exp(sum(BB3$coef\*Lev\_M)))

LA2\_m = LA2\*(exp(sum(BB2$coef\*Lev\_M)))

plot(p.time, LA3\_m, type = "s")

points(p.time, LA2\_m, type = "s", col = 2)

# 第三條線與第二條線的比值

plot(p.time, LA3\_m/LA2\_m, type = "s")

# 信賴區間

f.025=function(x){quantile(x, probs = 0.025, na.rm = FALSE,names = TRUE, type = 7)}

f.975=function(x){quantile(x, probs = 0.975, na.rm = FALSE,names = TRUE, type = 7)}

LA3\_bs = rbind(LA3, LA3\_m)

LA2\_bs = rbind(LA2, LA2\_m)

L23=ifelse(LA3\_bs/LA2\_bs=="NaN",0,LA3\_bs/LA2\_bs)

L23\_lower=apply(L23, 2, f.025)

L23\_uper=apply(L23, 2, f.975)

plot(p.time, L23\_uper, type="s",xlim=c(12,30))

points(p.time, L23\_lower, type="s")

abline(h=1, col=2)

# 結論：得到肝硬化是一種催化劑，會催化其他因子影響肝癌的程度