ACST3058

Yunbae Chae

2023-03-18

Contents

Abstract

I have highlighted detailed calculations where it seemed necessary, but I have mercilessly shortened calculations where it was obvious, like finding an expectation of an obvious $\sum_{i=1}^{n}(x_i - \mu)$. I just gave them 0 without explanations. Another example is the expectation of $\sum_{i=1}^{n}(x_i - \mu)^2$ as $n\sigma^2$. I hope you would understand and allow me to skip these sort of calculations quickly.

The intensity and cautions I have put in this assignment was for a potential full mark, so I would really appreciate if you could stay patient with my work until the end and mark them carefully. I am aware I have not calibrated well enough with the contents in Question 3 since the relevant tutorial solution material has not been uploaded. I have instead put much effort to research online and self-learn the Cox PH model materials. Please note that the order of completion of this assignment in terms questions was:

Question 1 -> Question 2 -> Question 4 -> Question 3.

So please do not be confused as you will notice my codes have become more streamlined for Question 4 than 3.

I hope you will enjoy and thank you for your time and effort for going through my work.

Question 1

Question 1 a

Using the relationship ${}_{s+t}p_x = {}_s p_x \cdot {}_t p_{x+s}$,

Consider ${}_2p_{x+1}$ first.

$$ {}_2p_{x+1} = {}_{1+1} p_{x+1} = {}_1 p_{x+1} \cdot {}_1 p_{x+1} = p_{x+1} \cdot p_{x+1} $$

Given that $p_{x+1} = 0.99$,

$$ p_{x+1} \cdot p_{x+1} = 0.99^2 $$

```
0.99^2
```

```
## [1] 0.9801
```

For ${}_3p_{x+1}$, it is known that ${}_2p_{x+1} = 0.99^2$ and $p_{x+2} = 0.985$

$$ {}_3p_{x+1} = {}_2 p_{x+1} \cdot {}_1 p_{x+2} = 0.99^2 + 0.985 $$

```
0.99^2+0.985
```

```
## [1] 1.9651
```

Question 1 b

Set up

To approximate the mean and variance of W, I will use the Delta method. Let W be in form $\frac{A}{B}$ where $A = Y_3,\ B = Y_1 + Y_2$:

$$g(\hat{\theta}) = g\left(\begin{bmatrix}\overline{A}\\ \overline{B}\end{bmatrix}\right) = \frac{\overline{A}}{\overline{B}}$$

The true value vector:

$$g(\overrightarrow{\mu}) = g\left(\begin{bmatrix}\mu_A\\ \mu_B\end{bmatrix}\right) = \frac{\mu_A}{\mu_B}$$

Please note that even when $B = Y_1 + Y_2$, Y_1 and Y_2 are independent (given that $Cov(Y_1, Y_2) = 0$) $\Rightarrow \mu_B = \mu_{Y_1} + \mu_{Y_2}$.

Next, I need the partial derivative vector:

$$\triangledown g(\overrightarrow{\mu}) = \begin{bmatrix}\frac{1}{\mu_A}\\ \frac{-\mu_B}{\mu_A^2}\end{bmatrix}$$

The variance-covariance matrix of the vector $\begin{bmatrix}\overline{A}\\ \overline{B}\end{bmatrix}$ using estimate of the covariance of the means from two samples ($Cov(\overline{A}, \overline{B}) = Cov(B, A)$):

$$\begin{bmatrix}\sigma_A^2/n & \sigma_{AB}\\ \sigma_{AB} & \sigma_B^2/n\end{bmatrix}$$

The Delta method is regarding the first two terms of the Taylor series:

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta)$$

Get variance first:

$$
\begin{aligned}
Var(g(\hat{\theta})) &\approx Var(g(\theta) + \triangledown g(\theta)^T \cdot (\hat{\theta} - \theta)) \\
&= Var(g(\theta) + \triangledown g(\theta)^T \cdot \hat{\theta} - \triangledown g(\theta)^T \cdot \theta) \\
&= Var(\triangledown g(\theta)^T \cdot \hat{\theta}) \\
&= \triangledown g(\theta)^T \cdot Cov(\hat{\theta}) \cdot \triangledown g(\theta) \\
&= \triangledown g(\theta)^T \begin{bmatrix} \sigma_A^2/n & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2/n \end{bmatrix} \triangledown g(\theta) \\
&= \frac{\sigma_A^2}{n\mu_B^2} - 2\frac{\sigma_{BA}\mu_A}{\mu_B^3} + \frac{\sigma_B^2\mu_A^2}{n\mu_B^4}
\end{aligned}
$$

Since $Cov[Y_1, Y_2] = Cov[Y_1, Y_3] = 0$ and $Cov[Y_2, Y_3] = 1$

$$\sigma_{AB} = \sigma_{Y_3Y_2} = 1,\ \sigma_B^2 = \sigma_{Y_1}^2 + \sigma_{Y_2}^2 = 1 + 2 = 3$$

$$\therefore Var(g(\hat{\theta})) = \frac{\sigma_A^2}{n\mu_B^2} - 2\frac{\mu_A}{\mu_B^3} + \frac{3\mu_A^2}{n\mu_B^4}$$

Conclusion

Mean and variance of $W = \frac{Y_3}{Y_1+Y_2}$ can be estimated by $g(\hat{\theta})$ with n=1:

$$g(\hat{\theta}) = g\left(\left[\frac{\overline{A}}{\overline{B}}\right]\right) = \frac{\overline{A}}{\overline{B}} = \frac{\frac{1}{n}(A_1 + \cdots A_n)}{\frac{1}{n}(B_1 + \cdots B_n)} = \frac{A_1}{B_1} = \frac{Y_3}{Y_1 + Y_2}$$

$$E(g(\hat{\theta})) \approx E(g(\overrightarrow{\mu})) + \triangledown g(\overrightarrow{\mu})^T E(\hat{\theta} - \overrightarrow{\mu})$$

$$\Rightarrow E\Big(\frac{Y_3}{Y_1 + Y_2}\Big) \approx E(\frac{\mu_A}{\mu_B}) + \triangledown g(\overrightarrow{\mu})^T E(\hat{\theta} - \overrightarrow{\mu}) = E(\frac{3}{3}) + \triangledown g(\overrightarrow{\mu})^T \cdot 0 = 1$$

$$Var(g(\hat{\theta})) = Var\Big(\frac{Y_3}{Y_1 + Y_2}\Big) \approx \frac{3}{1 \cdot 3^2} - 2\frac{3}{3^3} + \frac{3 \cdot 3^2}{1 \cdot 3^4} = \frac{4}{9}$$

Question 2

```
set.seed(1)
xdata <- rnorm(1000,mean=1,sd=1)
```

Question 2 a

At the start of part a, μ and σ are yet unknown and the number of samples (n) is also unknown.

I will start with a generic Normal function for $X \in \{X_1, X_2, \cdots, X_n\}$ and $\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\Big(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2 \Big)$$

$$L(\mu, \sigma^2; X) = (2\pi\sigma^2)^{-n/2} \cdot exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \Big)$$

$$lnL(\mu, \sigma^2; X) = -\frac{n}{2}ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$$

Calculations

$$\frac{\partial lnL(\mu, \sigma^2; X)}{\partial \mu} = 0 - \frac{\sum_{i=1}^{n} 2(x_i - \mu)\cdot -1}{2\sigma^2} = \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2}$$

$$\frac{\partial lnL(\mu, \sigma^2; X)}{\partial \sigma^2} = -\frac{n}{2}\cdot\frac{1}{\sigma^2} - \cdot -1 \cdot \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2(\sigma^2)^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2(\sigma^2)^2}$$

$$\frac{\partial^2 lnL(\mu, \sigma^2; X)}{\partial \mu^2} = \frac{\sum_{i=1}^{n}(0-1)}{\sigma^2} = -\frac{n}{\sigma^2}$$

$$\frac{\partial^2 lnL(\mu, \sigma^2; X)}{\partial \mu \partial \sigma^2} = -1 \cdot \frac{\sum_{i=1}^{n}(x_i-\mu)}{(\sigma^2)^2} = -\frac{\sum_{i=1}^{n}(x_i-\mu)}{(\sigma^2)^2}$$

$$\frac{\partial^2 lnL(\mu,\sigma^2;X)}{(\partial\sigma^2)^2} = -\frac{n}{2(\sigma^2)^2}\cdot -1 + -2\cdot\frac{\sum_{i=1}^n (x_i-\mu)^2}{2(\sigma^2)^3} = \frac{n}{2(\sigma^2)^2} - \frac{\sum_{i=1}^n (x_i-\mu)^2}{(\sigma^2)^3}$$

Results

$$\frac{\partial}{\partial\theta} lnL(\theta;X) = \begin{bmatrix} \frac{\sum_{i=1}^n (x_i-\mu)}{\sigma^2} \\ -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i-\mu)^2}{2(\sigma^2)^2} \end{bmatrix}$$

$$\frac{\partial^2}{\partial\theta\partial\theta^T} lnL(\theta;X) = \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{\sum_{i=1}^n (x_i-\mu)}{(\sigma^2)^2} \\ -\frac{\sum_{i=1}^n (x_i-\mu)}{(\sigma^2)^2} & \frac{n\sigma^2 - 2\sum_{i=1}^n (x_i-\mu)^2}{2(\sigma^2)^3} \end{bmatrix}$$

Writing the result functions into R:

```
mu <- 1
sigma2 <- 1

dmu <- function(xdata){
  sum(xdata-mu)/sigma2
}

dsigma2 <- function(xdata){
  -length(xdata)/(2*sigma2)+sum((xdata-mu)^2)/(2*sigma2^2)
}

dmu2 <- function(xdata){
  -length(xdata)/sigma2
}

dmudsigma2 <- function(xdata){
  -sum(xdata-mu)/sigma2^2
}

dsigma22 <- function(xdata){
  (length(xdata)*sigma2-2*sum((xdata-mu)^2))/(2*sigma2^3)
}

dtheta <- matrix(c(dmu(xdata),dsigma2(xdata)),2,1)
dtheta2 <- matrix(c(dmu2(xdata),dmudsigma2(xdata),dmudsigma2(xdata),dsigma22(xdata)),2,2)

dtheta
```

```
##           [,1]
```

```
## [1,] -11.64814
## [2,]  35.05771
```

```
dtheta2
```

```
##             [,1]       [,2]
## [1,] -1000.00000   11.64814
## [2,]    11.64814 -570.11542
```

The above output is the solution for

$$\frac{\partial}{\partial\theta} lnL(\theta; X) = \begin{bmatrix} \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2} \\ -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2(\sigma^2)^2} \end{bmatrix}$$

$$\frac{\partial^2}{\partial\theta\partial\theta^T} lnL(\theta; X) = \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{\sum_{i=1}^{n}(x_i-\mu)}{(\sigma^2)^2} \\ -\frac{\sum_{i=1}^{n}(x_i-\mu)}{(\sigma^2)^2} & \frac{n\sigma^2-2\sum_{i=1}^{n}(x_i-\mu)^2}{2(\sigma^2)^3} \end{bmatrix}$$

using true values $\mu = 1$, $\sigma = 1$, and the 1000 sampled xdata using the rnorm() function,

Whereas their expectations are:

$$E\Big(\frac{\partial}{\partial\theta} lnL(\theta; X)\Big) = \begin{bmatrix} 0 \\ -500 + \frac{n\sigma^2}{2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$E\Big(\frac{\partial^2}{\partial\theta^2} lnL(\theta; X)\Big) = \begin{bmatrix} -1000 & 0 \\ 0 & \frac{n\sigma^2-2n\sigma^2}{2(\sigma^2)^3} \end{bmatrix} = \begin{bmatrix} -1000 & 0 \\ 0 & -500 \end{bmatrix}$$

Next part should decide if these discrepancies were allowable.

Question 2 b

```
library(maxLik)
```

```
## Loading required package: miscTools
```

```
##
## Please cite the 'maxLik' package as:
## Henningsen, Arne and Toomet, Ott (2011). maxLik: A package for maximum likelihood estimation in R. C
##
## If you have questions, suggestions, or comments regarding the 'maxLik' package, please use a forum o
## https://r-forge.r-project.org/projects/maxlik/
```

```
normal.ll <- function(param){
mu <- param[1]
sd <- param[2]
logL <- -length(xdata)/2*log(2*pi*sd^2)-sum((xdata-mu)^2)/(2*sd^2)
logL
}
mle <- maxLik(normal.ll,start=c(mu=0.1,sigma=0.1))
print(summary(mle))
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 13 iterations
## Return code 8: successive function values within relative tolerance limit (reltol)
## Log-Likelihood: -1452.758
## 2  free parameters
## Estimates:
##       Estimate Std. error t value Pr(> t)
## mu     0.98835    0.03270   30.22  <2e-16 ***
## sigma  1.03440    0.02313   44.73  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

The estimates of μ and σ are therefore 0.98835 and 1.03440 respectively.

The conditions in this question should be a scenario of a 'large' sample size (1000), and therefore by the central limit theorem, $\hat{\theta} = (\theta_1, \theta_2)$ is assumed to be asymptotically normally distributed.

Then based on the lecture notes, an approximate $100(1-\alpha)\%$ confidence interval of θ_j is given by

$$\hat{\theta}_j \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\theta}_j)},\ j = 1, \cdots, k$$

$$^*\widehat{Var}(\hat{\theta}_j) = \widehat{Var}(\hat{\theta})_{jj}$$

Variance of MLE

The variance of the MLE $\hat{\theta}$ of θ can be estimated by using the Fisher's information matrix.

$$\widehat{Var}(\hat{\theta}) = \frac{1}{I(\hat{\theta})}$$

To get $I(\hat{\theta})$, get $I(\theta)$ first:

$$I(\theta) = -E\Big[\frac{\partial^2 lnL(\theta)}{\partial\theta\partial\theta^T}\Big] = -E\begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{\sum_{i=1}^n (x_i-\mu)}{(\sigma^2)^2} \\ -\frac{\sum_{i=1}^n (x_i-\mu)}{(\sigma^2)^2} & \frac{n\sigma^2 - 2\sum_{i=1}^n (x_i-\mu)^2}{2(\sigma^2)^3} \end{bmatrix}$$

$$= -\begin{bmatrix} -\frac{n}{\sigma^2} & 0 \\ 0 & -\frac{n}{2(\sigma^2)^2} \end{bmatrix}$$

$$I(\hat{\theta}) = \begin{bmatrix} \frac{1000}{\hat{\sigma}^2} & 0 \\ 0 & \frac{500}{(\hat{\sigma}^2)^2} \end{bmatrix}$$

$$I^{-1}(\hat{\theta}) = \widehat{Var}(\hat{\theta}) = \frac{1}{\frac{500000}{(\hat{\sigma}^2)^3} - 0\cdot 0}\begin{bmatrix} \frac{500}{(\hat{\sigma}^2)^2} & 0 \\ 0 & \frac{1000}{\hat{\sigma}^2} \end{bmatrix} = \begin{bmatrix} \frac{\hat{\sigma}^2}{1000} & 0 \\ 0 & \frac{\hat{\sigma}^2}{500} \end{bmatrix}$$

Using the estimates from R code to find $\hat{\theta}_j \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\theta}_j)},\ j = 1, 2$:

```
mu <- mle$estimate[[1]]
sigma <- mle$estimate[[2]]
theta <- matrix(c(mu,sigma),2,1)
V <- matrix(c(sigma^2/1000,sigma^2/500),2,1)
z <- qnorm(1-0.025)
CI <- matrix(c(theta-z*sqrt(V),theta+z*sqrt(V)),2,2)
CI
```

```
##           [,1]     [,2]
## [1,] 0.9242404 1.052463
## [2,] 0.9437309 1.125066
```

The R code return data is to be interpreted that μ has 95% CI within (0.9242404,1.0524633),

and σ within (0.9437309,1.1250656).

Question 3

Question 3 a

Below is the code to select relevant variables from the "burn" dataset:

```
library(KMsurv)
library(survival)
data(burn)
dat <- select(burn,Z2,Z3,Z4,T1,D1)
```

Using coxph() function, the covariates can be fitted with the time to excision and censorship status, then using cox.zph(), I will get the p-values to see if it is a good fit:

```
fit <- coxph(Surv(T1,D1)~Z2+Z3+Z4, data=dat)
summary(fit)
```

```
## Call:
## coxph(formula = Surv(T1, D1) ~ Z2 + Z3 + Z4, data = dat)
##
##   n= 154, number of events= 99
##
##         coef exp(coef)  se(coef)      z Pr(>|z|)
## Z2  0.652004  1.919383  0.228757  2.850  0.00437 **
## Z3  0.142730  1.153419  0.303015  0.471  0.63762
## Z4 -0.006345  0.993675  0.005546 -1.144  0.25257
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##    exp(coef) exp(-coef) lower .95 upper .95
## Z2    1.9194      0.521    1.2259     3.005
## Z3    1.1534      0.867    0.6369     2.089
## Z4    0.9937      1.006    0.9829     1.005
##
## Concordance= 0.57  (se = 0.033 )
## Likelihood ratio test= 8.38  on 3 df,   p=0.04
## Wald test            = 8.99  on 3 df,   p=0.03
## Score (logrank) test = 9.23  on 3 df,   p=0.03
```

The dat$Z2 covariate is the 'gender' covariate. With its low p-value (<0.05), the gender covariate has a statistically significant effect on the time to excision. The rest of the covariates, 'race' and 'percentage of burned surface' have no statistically significant effects on the time to excision. In everyday words, female patients have a statistically significantly shorter time to excision, than male patients.

Measuring each of the covariate's effect being constant over time would indicate whether the test is a good fit and there is credibility in the statement I just made about the covariates.

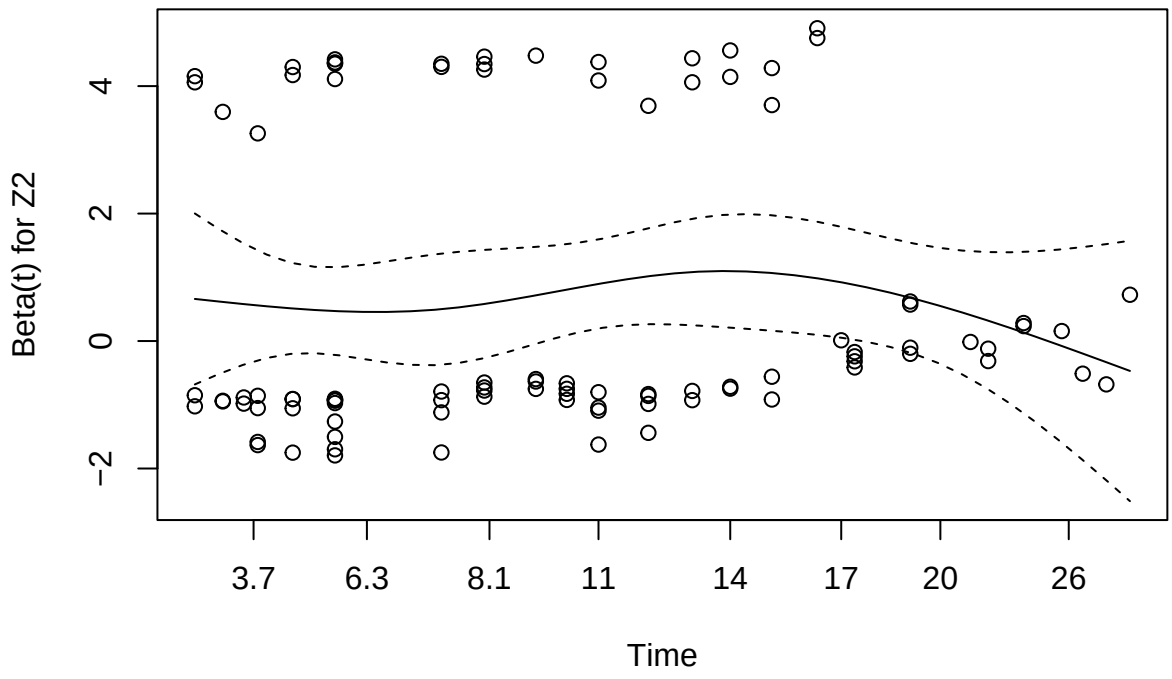
```
cox.zph(fit)
```

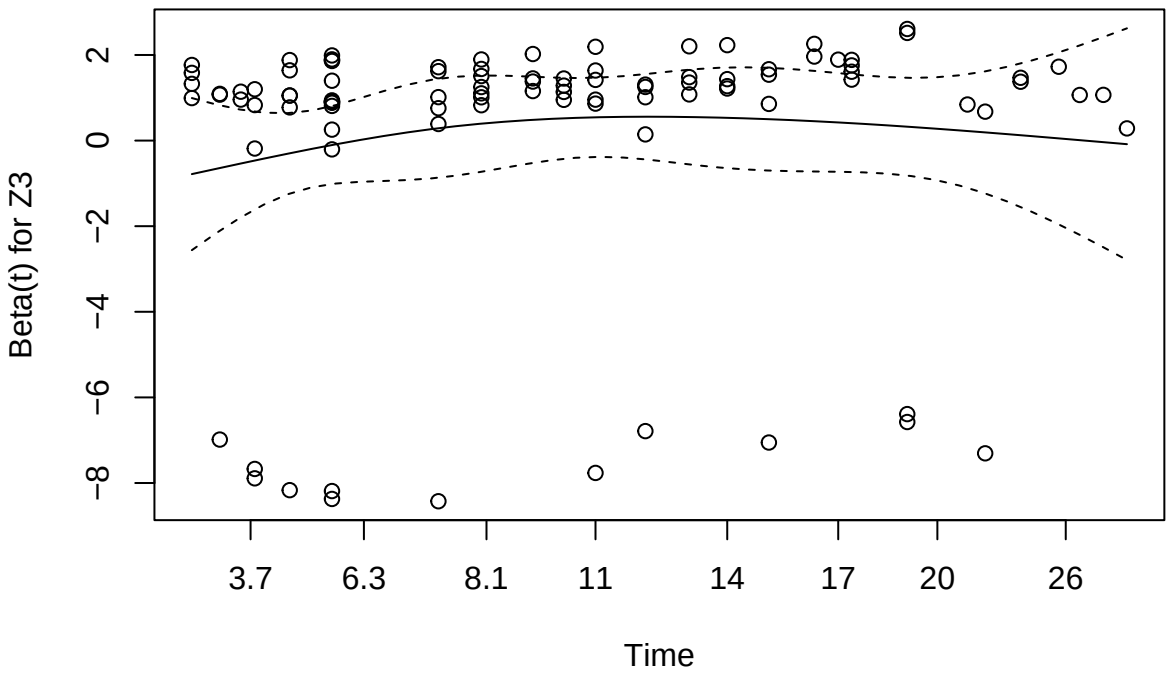
```
##           chisq df    p
## Z2     2.97e-05  1 1.00
## Z3     9.55e-01  1 0.33
## Z4     1.94e+00  1 0.16
## GLOBAL 2.83e+00  3 0.42
```

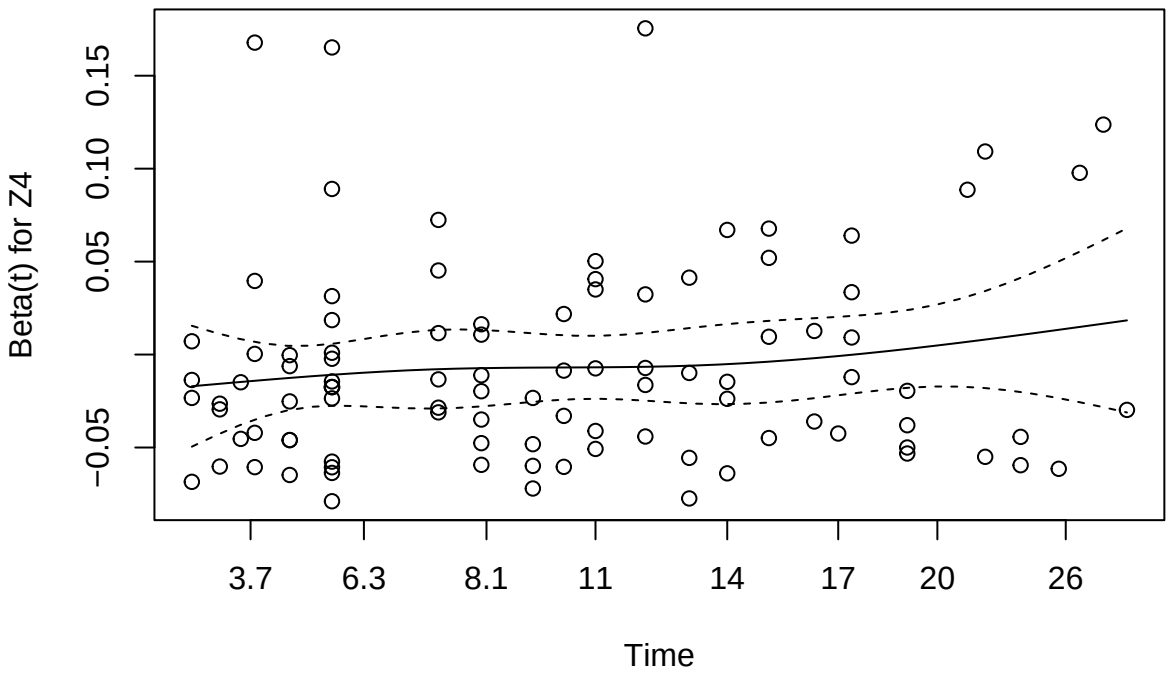
All of the 4 p-values including that of the GLOBAL (weighted average of each covariates) being higher than 0.05 favours H_0: there is constant effect of covariates over time (random dispersion of residuals around 0). H_0 is not rejected and therefore the proportional hazards assumption holds.

To visualise this:

```
plot(cox.zph(fit))
```




Beta(t) for Z3
2
0
−2
−4
−6
−8
3.7
6.3
8.1
11
14
17
20
26
Time



Along with those p-values, it can be double confirmed that the scaled Schoenfeld residuals trend around 0 with good randomness.

It is a well-fitting model.

Question 3 b

The survfit() function allows me to create a new survival curve for the specific covariate arguments that I give. I will give Z2 gender = 0 male, Z3 race = 1 white, and Z4 % area burned = 50:

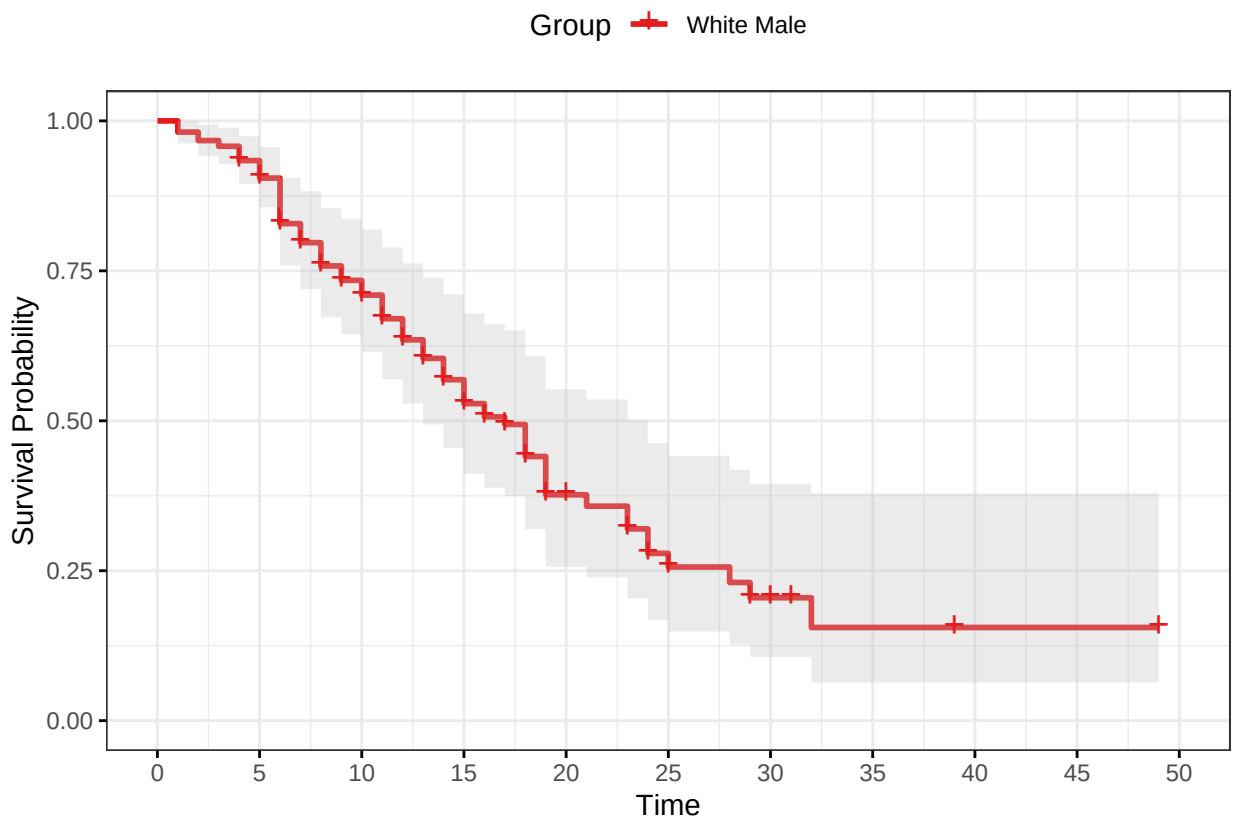
```
library(survminer)
```

```
## Loading required package: ggpubr
```

```
##
## Attaching package: 'survminer'
```

```
## The following object is masked from 'package:survival':
##
##     myeloma
```

```
library(ggplot2)
newdat <- data.frame(Z2=0,Z3=1,Z4=50)
newcurve <- survfit(fit,newdata=newdat)
ggsurvplot(newcurve, conf.int = 0.95, data=burn,
           break.time.by = 5,
           ggtheme = theme_bw(),
           palette = "Set1",
           legend.labs = c("White Male"),
           legend.title = "Group",
           xlab = "Time",
           ylab = "Survival Probability",
           main = "Survival Curve with 95% Confidence Interval for White Male")
```



To find $\hat{S}(30)$, I can use the summary() function:

```
summary(newcurve,times=30)
```

```
## Call: survfit(formula = fit, newdata = newdat)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    30      5      98    0.205  0.0686        0.106        0.395
```

The estimated $\hat{S}(30)$ is 0.2049641.

Question 3 c

The model being the Cox proportional hazards model, the coefficients for each of the three variables are log hazard ratios.

```
summary(fit)[7]
```

```
## $coefficients
##            coef exp(coef)    se(coef)          z    Pr(>|z|)
## Z2  0.652003619  1.919383 0.228756552  2.8502074 0.004369073
## Z3  0.142730317  1.153419 0.303015219  0.4710335 0.637616818
## Z4 -0.006345117  0.993675 0.005545817 -1.1441266 0.252571145
```

The coefficient 0.652003619 for gender means the log hazard of time to excision is 0.652003619 units higher for females (R2=1) than males (R2=0), holding other variables constant. The hazard rate for females is $e^{0.652003619}$ times higher than the hazard rate for males ($\approx$ 1.9193827 times higher!). Therefore, the data should show significantly shorter time to excision for females than males.

Similarly, for race, if the coefficient for race is 0.142730317, the log hazard of time to excision is 0.142730317 units higher for white patients than for nonwhite patients ($\approx$ 1.1534187 times higher), all other variables held constant.

For the % total area burned, one unit increase in the percentage of total surface area burned decreases 0.006345117 unit in the log hazard of time to excision ($\approx$ 0.6325029% decrease in time to excision). The larger the surface, the shorter the time to excision, all else being equal.

Throughout the earlier parts of the question, the Cox PH model was used to determine gender as the only statistically significant covariate. Upon examining the coefficients, it seems to match up because other coefficients than gender are not high enough to be considered meaningfully effective.

Question 4

After carefully inspecting d1, d2 and d3 specifications, it was clear that the data has no information about truncation and only d1 is relevant for the question. All patients who relapsed died and their death times were all included in the data. So ultimately, this data contains only death or censorship times. All living surviving patients had t1 equal to t2. In this way, I only needed columns t1 and d1.

I will generate a table for $\hat{S}(t)$ and $\hat{\Lambda}(t)$ and compare it with the result using the survfit() function. *I attempted this question prior to attempting question 3. After answering everything in this question, I learned that generating the table like below is an unnecessary labor that can be easily replaced by the survfit() function. The generation of the table is still included just to highlight matching results from the two methods, which confirms that I have had a correct understanding in the scope of the data and how the variables relate to each other.

I used the following definitions for $\hat{S}(t)$ and $\hat{\Lambda}(t)$:

$$\hat{S}(t_i) = \hat{S}(t_{i-1})\Big(1 - \frac{d_2}{r_2}\Big),\ \hat{S}(t_0) = 1$$

$$\hat{\Lambda}(t_j) = \hat{\Lambda}(t_{j-1} + \frac{d_j}{r_j}),\ \hat{\Lambda}(t_0) = 0$$

```
data(bmt)

dat <- filter(bmt,group==1)
dat <- select(dat,t1,d1)
dat <- arrange(dat,t1)

Gen <- function(da){
  tab <- data.frame(i=numeric(),ti = numeric(),d = numeric(),c = numeric(),r = numeric(),S = numeric(),
  newrow <- 0
  d <- 0
  ii <- 0
  ti <- 0
  c <- 0
  r <- nrow(da)
  S <- 1
  N <- 0
  j <- 1
  end <- nrow(da)+1
  for(i in 1:nrow(da)){
    if(da[i,1]!=da[i-1,1]||i==1){
      S <- S*(1-d/r)
      N <- N+d/r
      tab[j,] <- c(ii,ti,d,c,r,S,N)
      j <- j+1
      ii <- ii+1
      ti <- da[i,1]
```

```
      r <- r-d-c
      d <- 0
      c <- 0
      if(da[i,2]==1){
        d <- 1
      }else{
        c <- 1
      }
    }else{
      if(da[i,2]==1){
        d <- d+1
      }else{
        c <- c+1
      }
    }
  }
  tab
}

Gen(dat)
```

```
##     i   ti d c  r         S          N
## 1   0    0 0 0 38 1.0000000 0.00000000
## 2   1    1 1 0 38 0.9736842 0.02631579
## 3   2   86 1 0 37 0.9473684 0.05334282
## 4   3  107 1 0 36 0.9210526 0.08112059
## 5   4  110 1 0 35 0.8947368 0.10969202
## 6   5  122 1 0 34 0.8684211 0.13910379
## 7   6  156 1 0 33 0.8421053 0.16940682
## 8   7  162 1 0 32 0.8157895 0.20065682
## 9   8  172 1 0 31 0.7894737 0.23291488
## 10  9  194 1 0 30 0.7631579 0.26624822
## 11 10  226 0 1 29 0.7631579 0.26624822
## 12 11  243 1 0 28 0.7359023 0.30196250
## 13 12  262 2 0 27 0.6813910 0.37603658
## 14 13  269 1 0 25 0.6541353 0.41603658
## 15 14  276 1 0 24 0.6268797 0.45770324
## 16 15  350 1 0 23 0.5996241 0.50118150
## 17 16  371 1 0 22 0.5723684 0.54663605
## 18 17  417 1 0 21 0.5451128 0.59425510
## 19 18  418 1 0 20 0.5178571 0.64425510
## 20 19  466 1 0 19 0.4906015 0.69688668
## 21 20  487 1 0 18 0.4633459 0.75244223
## 22 21  526 1 0 17 0.4360902 0.81126576
## 23 22  530 0 1 16 0.4360902 0.81126576
## 24 23  716 1 0 15 0.4070175 0.87793243
## 25 24  781 1 0 14 0.3779449 0.94936100
## 26 25  996 0 1 13 0.3779449 0.94936100
## 27 26 1111 0 1 12 0.3779449 0.94936100
## 28 27 1167 0 1 11 0.3779449 0.94936100
## 29 28 1182 0 1 10 0.3779449 0.94936100
## 30 29 1199 0 1  9 0.3779449 0.94936100
## 31 30 1279 1 0  8 0.3307018 1.07436100
```

```
## 32 31 1330 0 1  7 0.3307018 1.07436100
## 33 32 1377 0 1  6 0.3307018 1.07436100
## 34 33 1433 0 1  5 0.3307018 1.07436100
## 35 34 1462 0 1  4 0.3307018 1.07436100
## 36 35 1496 0 1  3 0.3307018 1.07436100
## 37 36 1602 0 1  2 0.3307018 1.07436100
```

Question 4 a

$$\hat{S}(500) = \hat{S}(487)$$

In the above table, $\hat{S}(487) = 0.4633459$:

```
select(filter(Gen(dat),ti==487),S)
```

```
##           S
## 1 0.4633459
```

Now there is the survival package:

```
survdat <- bmt[bmt$group==1,]
fit <- survfit(Surv(survdat$t1,survdat$d1)~1,data=survdat)
S <- summary(fit,times=500)$surv[1]
S
```

```
## [1] 0.4633459
```

$$\therefore \hat{S}(500) = \hat{S}(487) = 0.4633459$$

Question 4 b

For NAE, I used cumsum() function:

```
NAE <- cumsum(fit$n.event/fit$n.risk)[20]
NAE
```

```
## [1] 0.7524422
```

```
select(filter(Gen(dat),ti==487),N)
```

```
##           N
## 1 0.7524422
```

$$\hat{S}(500) = \hat{S}(487) = e^{-\hat{\Lambda}(487)} = e^{-0.7524422} = 0.4712143$$

Question 4 c

```
dat <- filter(bmt,group==c(2,3))
```

```
## Warning: There was 1 warning in `filter()`.
## i In argument: `group == c(2, 3)`.
## Caused by warning in `group == c(2, 3)`:
## ! longer object length is not a multiple of shorter object length
```

```
dat <- select(dat,group,t1,d1)
for(i in 1:nrow(dat)){
  if(dat[i,1]==2){
    dat[i,1] <- "AML Low Risk"
  }else{
    dat[i,1] <- "AML High Risk"
  }
}
fitdat <- survfit(Surv(dat$t1,dat$d1)~dat$group,data=dat,conf.type="plain",conf.int=0.95)
ggsurvplot(fitdat, conf.int = 0.95, censor= F,
           ggtheme = theme_minimal())
```