BIOST 537 HOMEWORK1

Ivy Zhang

1/15/2022

Problem 1

(a) Individual A

It is right censored. Because the individual A does not have event before the study ends. It is also a left truncated. Because if individual A had breast cancer diagnosis before age 32, she won't be enrolled.

(b) Individual B

It is interval censorded. The individual B has the event in the interval three years between the fourth exam and the fifth exam. It is also a left truncated. Because if individual B had breast cancer diagnosis before age 39, she won't be enrolled.

(c) Individual C

It is right censored. The researchers cannot observe whether the individual C has event or not anymore because of other random reasons that are not related to the event. It is also a left truncated. Because if individual C had a breast cancer diagnosis before age 55, she won't be enrolled into this study.

(d) Individual D

It is right censored. The researchers cannot observe whether the individual D has event or not anymore because of other random reasons that are not related to the event. It is also a left truncated. Because if individual D had a breast cancer diagnosis before age 45, she won't be enrolled into this study.

Interested in studying the time from enrollment instead of age at onset

For all of the previous four cases, there won't be left truncated any more, but they will be still censored.

Problem 2

(a) Truncation: Some people who have extremely short time between the diagnosis and first bowel resection surgery or death before the enrollment date or having bowel resection surgery or death before the diagnosis may be left truncated from the study. However, people who may be diagnosed earlier than previously mentioned population but have longer time between diagnosis and first bowel resection surgery or death will be enrolled into the study, and people who have diagnosis before death and first bowel resection surgery can be enrolled. It may lead to overestimation.

Censoring: People who have dropped out from the study because of some reasons before they have first bowel resection surgery or death will be censored and people who do not have first bowel resection surgery or death in ten years are censored.

(b) No. Because we are enrolling people who only have diagnosis but do not have the first bowel resection surgery or death, people with older diagnosis age may be more likely to have first bowel resection surgery or death, meaning they may be less likely to be enrolled in the study. We will underestimate the diagnosis age.

Problem 3

(a)Study population

Study population consistude of two parts. One part is 312 primary biliary cirrhosis patients who attended esither of two double-blind, placebo-controlled, randomized clinical trials at the Mayo clinic. All of them meet well-estibalished clinical, biochemical, serologic and histologic criteria for primary biliary cirrhosis. Their data were used for model development. Another part is the 112 patients who were eligible for the trials but declined to participate, their data were used for model validation.

(b)Initiating event

The initiating event is to be determined of eligibile for the trials.

(c)Terminating event

The terminating event is death from any cause.

(d)Time scale

The time scale is months from the participants entering the trail to censoring or terminating event.

(e)Causes of censoring

The cause of censoring maybe the liver transplantation, lost to follow-up, or do not have death after the trial ends.

(f)Comment on whether you believe the underlying censoring mechanism may be related to the outcome of interest?

If we assume the reasons why people lost to follow-up are not related to their death, then it seems to be safe to assume it is uninformative censoring. Same for the lost to follow-up, if we assume the reasons of the participants lost to follow-up is not related to their death, then it is also safe to assume it is uninformative censoring. However, there are also some cases that people who lost to follow-up is because they are too ill to respond or they are relatively healthy to relocate, then it will be risky to assume it is uninformative censoring in these cases. For liver transplantation, it may be the case that people who are more likely to die will have higher possibility to have liver transplantation, then it may be very risky to assume it is uninformative censoring.

(g)Five year survival rate

Based on the figure 3 Kaplan-Meier survival curve, the low-risk group has around 90% of the participants survive within five years, meaning 10% of the participants in the low-risk group dies within five years, the medium-risk group has 50%, meaning around 50% of the participants in middle-risk group die within 5 year, and the high-risk group has 20%, meaning around 80% of the participants in the high-risk group die in five years.

Problem 4

(a)Compute the average follow-up time and the proportion of censored observations.

```
library(readr)
addicts <- read_csv("addicts.csv")
mean(addicts$time)
```

```
## [1] 402.5714
```

```
1-mean(addicts$event)
```

```
## [1] 0.3697479
```

Based on the calculation, the average follow-up time is 402.57 days, and 0.37 of the data are censored.

(b) fit exponential, Weibull and generalized gamma models

Fitting exponential model and report parameter estimates, 95% confidence intervals, and maximum loglilihood value:

```
library(survival)
library(foreign)
library(flexsurv)
library(msm)
library(knitr)
s.addicts = with(addicts, Surv(time, event))
source("fitparametric.R")
addicts.exp = suppressMessages(fitparametric(s.addicts, dist="exp"))
exp.tab = matrix(round(addicts.exp$coeff,4)[,-4])
exp.tab = t(exp.tab)
exp.tab = cbind(exp.tab, addicts.exp$loglik)
colnames(exp.tab) = c("Estimate","95% confindence interval lower bound",
                      "95% confindence interval lower bound",
                       "Maximum Likelihood")
rownames(exp.tab) = "lambda"
```

```
kable(exp.tab)
```

	Estimate	95% confindence interval lower bound	95% confindence interval lower bound	Maximum Likelihood
lambda	0.0016	0.0013	0.0018	-1118.93

Fitting Weibull model and report parameter estimates, 95% confidence intervals, and maximum log likelihood

value:

```
addicts.weib = fitparametric(s.addicts, dist="weibull")
weib.tab = matrix(round(addicts.weib$coeff,4)[,-4], nrow = 2)
weib.tab = cbind(weib.tab, addicts.weib$loglik)
colnames(weib.tab) = c("Estimate","95% confindence interval lower bound",
                       "95% confindence interval lower bound",
                       "Maximum Likelihood")
rownames(weib.tab) = c("lambda","p")
```

```
kable(weib.tab)
```

	Estimate	95% confindence interval lower bound	95% confindence interval lower bound	Maximum Likelihood
lambda	0.0016	0.0014	0.0018	-1114.92
p	1.2264	1.0603	1.3925	-1114.92

Fitting generalized gamma model and report parameter estimates, 95% confidence intervals, and maximum loglilihood value:

```
addicts.gengamma = fitparametric(s.addicts, dist="gengamma")
gengamma.tab = matrix(round(addicts.gengamma$coeff,4)[,-4], nrow = 3)
gengamma.tab = cbind(gengamma.tab, addicts.gengamma$loglik)
colnames(gengamma.tab) = c("Estimate","95% confindence interval lower bound",
                           "95% confindence interval lower bound",
                         "Maximum Likelihood")
rownames(gengamma.tab) = c("mu","sigma","Q")
```

```
kable(gengamma.tab)
```

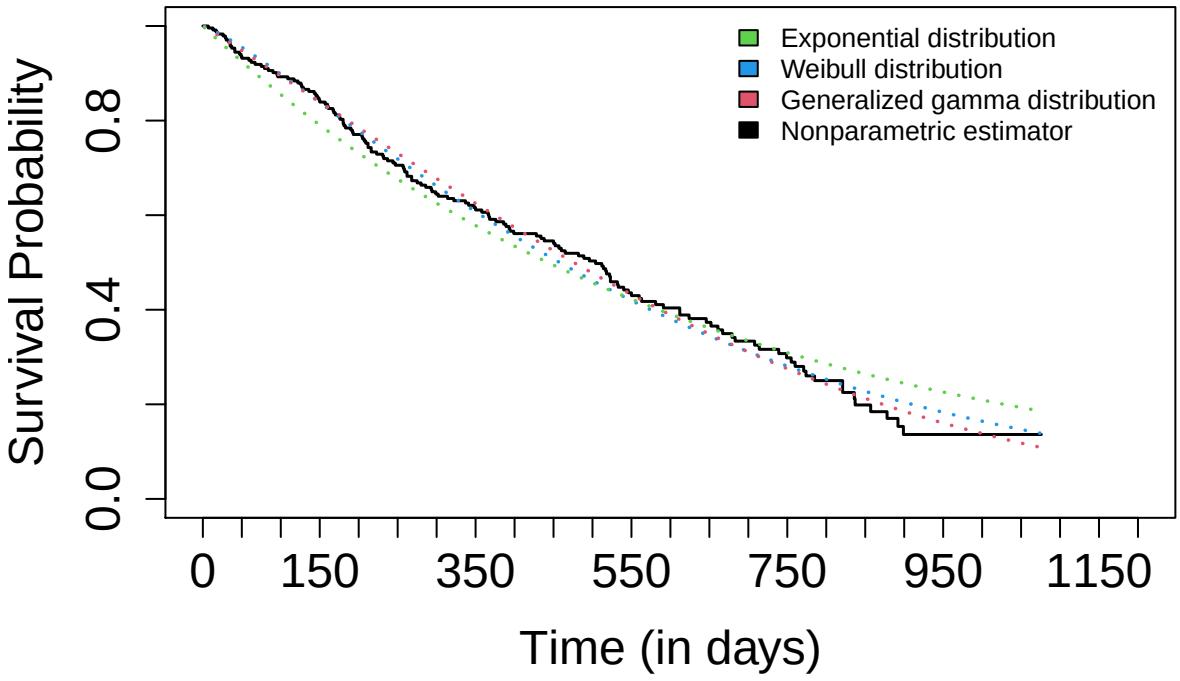
	Estimate	95% confindence interval lower bound	95% confindence interval lower bound	Maximum Likelihood
mu	6.5502	6.2694	6.8311	-1114.36
sigma	0.6595	0.3260	0.9930	-1114.36
Q	1.4682	0.3848	2.5516	-1114.36

(c) Plot the survival function

```
addicts.fitweibull <- flexsurvreg(s.addicts ~
    1, data = addicts, dist = "weibull")
addicts.fitggamma <- flexsurvreg(s.addicts ~
    1, data = addicts, dist = "gengamma")
addicts.fitexp <- flexsurvreg(s.addicts ~
    1, data = addicts, dist = "exp")

plot(survfit(s.addicts ~ 1, data = addicts),
    conf.int = FALSE, mark.time = FALSE, xaxt = "n", cex.axis = 1.5,
    cex.lab = 1.5,xlim = c(0,1200),
    xlab = "Time (in days)", ylab = "Survival Probability",
    lwd = 1.5)
axis(1, at = seq(0, 1200, by = 50), labels = seq(0,
```

```
    1200, by = 50), cex.axis = 1.5)
lines(addicts.fitweibull, col = 4, ci = FALSE, lwd = 1.8,
    lty = 3)
lines(addicts.fitggamma, col = 2, ci = FALSE, lwd = 1.8,
    lty = 3)
lines(addicts.fitexp, col = 3, ci = FALSE, lwd = 1.8,
    lty = 3)
legend("topright", legend = c("Exponential distribution",
                              "Weibull distribution",
                              "Generalized gamma distribution ",
    "Nonparametric estimator "), fill = c(3,4,2,1 ), cex = 0.8, bty = "n")
```



Although the three parametric model looks all fit well with the non-parametric model, it seems that generalized gamma distribution model fits best to the non-parametric model compared to the rest two models.

(d) Is the Weibull model an appropriate simplication of the generalized gamma model in this example?

```
2 * (addicts.gengamma$loglik - addicts.weib$loglik)
```

```
## [1] 1.12
```

```
1 - pchisq(2 * (addicts.gengamma$loglik - addicts.weib$loglik),
    df = 1)
```

```
## [1] 0.2899185
```

We used the log maximum likelihood of the two fitted model to perform the LRT test. Our test statistics is around 1.12. We calculatedthe p-value gave by the lrt test, which is around 0.290 that is larger than 0.05. We conclude that the we fail to reject the null hypothesis that the Weibull model is an appropriate simplification of the generalized gamma model at the significance level of 0.05.

(e) Using a Weibull model, provide an estimate and 95% confidence interval of:

i. the median time until exit from maintenance;

```
library(numDeriv)
median_tab = fitparametric(s.addicts, dist="weibull",feature="quantile")
```

```
kable(round(median_tab$feature[,-4],3))
```

	x
estimate	457.739
ci.lower	396.989
ci.upper	518.490

We estimate that the median time until exit from maintenance is 457.739 (95%CI: 396.989, 518.490) days using Weibull model.

ii. the probability that no exit will occur by one year;

```
year_tab = fitparametric(s.addicts, dist="weibull",feature="survival",t=365)
```

```
kable(round(year_tab$feature[,-4],3))
```

	x
estimate	0.591
ci.lower	0.538
ci.upper	0.645

We estimate that the probability that no exit will occur by one year is 0.592 (95%CI: 0.538, 0.645) using Weibull model.

iii. the probability that no exit will occur by two years given that no exit has occurred by one year.

```
cons_tab = fitparametric(s.addicts, dist="weibull",feature="condsurvival",t=730, t0 = 365)
```

```
kable(round(cons_tab$feature[,-4],3))
```

	x
estimate	0.495
ci.lower	0.422
ci.upper	0.567

We estimate that the probability that no exit will occur by two years given that no exit has occurred by one year is 0.495 (95%CI: 0.422, 0.567) using Weibull model.

(f) Is the exponential model an appropriate simplification of the Weibull model in this example?

```
2 * (addicts.weib$loglik - addicts.exp$loglik)
```

```
## [1] 8.02
```

```
1 - pchisq(2 * (addicts.weib$loglik - addicts.exp$loglik),
    df = 1)
```

```
## [1] 0.004626357
```

We used the log maximum likelihood of the two fitted model to perform the LRT test. Our statistics is calculated as 8.02. Based on the p-value gave by the LRT, which is around 0.005 that is larger than 0.05. We conclude that the we have evidence from this data to support that the exponential model is not an appropriate simplification of the Weibull model in this example at the significance level of 0.05.

(g) Separately fit an exponential model to the subset of individuals in clinic 1 and clinic 2. Report parameter estimates and corresponding 95% confidence intervals.

```
fit.exp.clinic1 = flexsurvreg(Surv(time, event) ~1, data = addicts[addicts$clinic == 1,],
                              dist = "exponential")

clinic_tab = matrix(round(fit.exp.clinic1$res,5)[,-4],byrow = T,nrow = 1)
fit.exp.clinic2 = flexsurvreg(Surv(time, event) ~1, data = addicts[addicts$clinic == 2,],
                              dist = "exponential")
clinic_tab = rbind(clinic_tab,round(fit.exp.clinic2$res,5)[,-4])
colnames(clinic_tab) = c("Estimate","95%CI lower","95%CI upper")
rownames(clinic_tab) = c("Clinic 1 lambda","Clinic 2 lambda")
kable(clinic_tab)
```

	Estimate	95%CI lower	95%CI upper
Clinic 1 lambda	0.00205	0.00172	0.00245
Clinic 2 lambda	0.00077	0.00053	0.00112

```
delta = fit.exp.clinic1$res[1] - fit.exp.clinic2$res[1]
delta_se = sqrt(fit.exp.clinic1$res[4]^2 + fit.exp.clinic2$res[4]^2)
T_W = abs(delta) / delta_se
2 * pnorm(-T_W)
```

```
## [1] 6.404186e-08
```

We assume the two sub-populations are independent, and calculate the standard error and difference between the lambda estimation based on the two previous fitted model. Then we use the Wald test and calculated two-sided p-value. As we can see, the two-sided p-value is much smaller than 0.05. Thus, we conclude that we reject the null hypothesis that the distribution of time exit from maintenance is same in the two clinics at the significance level of 0.05.

(h) Repeat the last problem but substituting clinic by history of incarceration (i.e., prison).

```
fit.exp.prison1 = flexsurvreg(Surv(time, event) ~1, data = addicts[addicts$prison == 1,],
                              dist = "exponential")

prison_tab = matrix(round(fit.exp.prison1$res,5)[,-4],byrow = T,nrow = 1)
fit.exp.prison = flexsurvreg(Surv(time, event) ~1, data = addicts[addicts$prison == 0,],
                              dist = "exponential")
prison_tab = rbind(prison_tab,round(fit.exp.prison$res,5)[,-4])
colnames(prison_tab) = c("Estimate","95%CI lower","95%CI upper")
```

```
rownames(prison_tab) = c("Prison lambda","Non-Prison lambda")
kable(prison_tab)
```

	Estimate	95%CI lower	95%CI upper
Prison lambda	0.00172	0.00136	0.00218
Non-Prison lambda	0.00145	0.00117	0.00181

```
delta = fit.exp.prison1$res[1] - fit.exp.prison$res[1]
delta_se = sqrt(fit.exp.prison1$res[4]^2 + fit.exp.prison$res[4]^2)
T_W = abs(delta) / delta_se
2 * pnorm(-T_W)
```

```
## [1] 0.3064417
```

We assume the two sub-populations are independent, and calculate the standard error and difference between the lambda estimation based on the two previous fitted model. Then we use the Wald test and calculated two-sided p-value. As we can see, the two-sided p-value is 0.306 which is larger than 0.05. Thus, we conclude that we cannot reject the null hypothesis that the distribution of time exit from maintenance is same for people who have different history of incarceration at the significance level of 0.05.