100/100

# Survival Analysis Computing HW

[redacted]

25 October 2022

## 1

(a) Given the failure time $T \sim$Exp$(\theta)$ and the censor time $C$ has some other distribution, such that the observed data is $Y = \min(T, C)$ and the indicator $\Delta = I(T \leq C)$, one can write the likelihood using methods presented in chapter 2, given the assumption of general, not conditional, independence between each pair/instance of $C$ and $T$. It is also given in the problem statement that the observed data are i.i.d., such that observations are independent across individuals and participants within each of the two study arms have identical hazard rates, $\theta_1$ for the placebo arm and $\theta_2$ for the treatment arm.

Under the above assumptions, the likelihood $\mathcal{L}$ can be written as follows, where $f(t;\theta)$, $S(t;\theta)$ are the density and survival functions of the failure time $T$ and $g(t), G(t)$ are the corresponding functions for the censor time $C$. The below uses the notation given in the problem statement.

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} \{[f(x_{1i};\theta_1)G(x_{1i})]^{\delta_{1i}}[S(x_{1i};\theta_1)g(x_{1i})]^{1-\delta_{1i}}\} \times \\
&\quad \prod_{j=1}^{m} \{[f(x_{2j};\theta_2)G(x_{2j})]^{\delta_{2j}}[S(x_{2j};\theta_2)g(x_{2j})]^{1-\delta_{2j}}\} \\
&= \prod_{i=1}^{n} \{[f(x_{1i};\theta_1)^{\delta_{1i}} S(x_{1i};\theta_1)^{1-\delta_{1i}}][g(x_{1i})^{1-\delta_{1i}} G(x_{1i})^{\delta_{1i}}]\} \times \\
&\quad \prod_{j=1}^{m} \{[f(x_{2j};\theta_2)^{\delta_{2j}} S(x_{2j};\theta_2)^{1-\delta_{2j}}][g(x_{2j})^{1-\delta_{2j}} G(x_{2j})^{\delta_{2j}}]\}
\end{aligned}$$

Taking the above form and plugging in the known forms of the Exponential density $f(t;\theta) = \theta e^{-\theta t}$ and corresponding survival function $S(t;\theta) = e^{-\theta t}$, one finds the following likelihood function.

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n}\{[f(x_{1i};\theta_1)^{\delta_{1i}}S(x_{1i};\theta_1)^{1-\delta_{1i}}][g(x_{1i})^{1-\delta_{1i}}G(x_{1i})^{\delta_{1i}}]\}\times \\
&\quad \prod_{j=1}^{m}\{[f(x_{2j};\theta_2)^{\delta_{2j}}S(x_{2j};\theta_2)^{1-\delta_{2j}}][g(x_{2j})^{1-\delta_{2j}}G(x_{2j})^{\delta_{2j}}]\} \\
&= \prod_{i=1}^{n}\{[\theta_1^{\delta_{1i}}e^{-\theta_1 x_{1i}\delta_{1i}}e^{-\theta_1 x_{1i}(1-\delta_{1i})}][g(x_{1i})^{1-\delta_{1i}}G(x_{1i})^{\delta_{1i}}]\}\times \\
&\quad \prod_{j=1}^{m}\{[\theta_2^{\delta_{2j}}e^{-\theta_2 x_{2j}\delta_{2j}}e^{-\theta_2 x_{2j}(1-\delta_{2j})}][g(x_{2j})^{1-\delta_{2j}}G(x_{2j})^{\delta_{2j}}]\} \\
&= \prod_{i=1}^{n}\{[\theta_1^{\delta_{1i}}e^{-\theta_1 x_{1i}}][g(x_{1i})^{1-\delta_{1i}}G(x_{1i})^{\delta_{1i}}]\}\times \\
&\quad \prod_{j=1}^{m}\{[\theta_2^{\delta_{2j}}e^{-\theta_2 x_{2j}}][g(x_{2j})^{1-\delta_{2j}}G(x_{2j})^{\delta_{2j}}]\}
\end{aligned}
$$

Without further knowledge concerning the censoring time distribution, one cannot simplify the full likelihood further. However, as this likelihood will be maximized over only the parameters $\theta_1, \theta_2$, it can be written as the relevant portion containing these parameters and a proportionality constant. This is undergone below, after the full likelihood is presented as the final answer.

$$
\boxed{\mathcal{L} = \prod_{i=1}^{n}\{[\theta_1^{\delta_{1i}}e^{-\theta_1 x_{1i}}][g(x_{1i})^{1-\delta_{1i}}G(x_{1i})^{\delta_{1i}}]\}\prod_{j=1}^{m}\{[\theta_2^{\delta_{2j}}e^{-\theta_2 x_{2j}}][g(x_{2j})^{1-\delta_{2j}}G(x_{2j})^{\delta_{2j}}]\}}
$$

$$
\implies \mathcal{L} = \prod_{i=1}^{n}[\theta_1^{\delta_{1i}}e^{-\theta_1 x_{1i}}]\prod_{j=1}^{m}[\theta_2^{\delta_{2j}}e^{-\theta_2 x_{2j}}] \times c
$$

$$
\implies \mathcal{L} \propto \prod_{i=1}^{n}[\theta_1^{\delta_{1i}}e^{-\theta_1 x_{1i}}]\prod_{j=1}^{m}[\theta_2^{\delta_{2j}}e^{-\theta_2 x_{2j}}] = L
$$

(b) To derive the MLE, the likelihood presented above must be maximized over the distribution parameters $\theta_1, \theta_2$. This proceeds using the factor of the likelihood which contains the $\theta$ parameters of interest. This can be done because those factors in the likelihood which do not depend on the parameters $\theta = \{\theta_1, \theta_2\}$ can be considered a proportionality constant here and disregarded. This is due to the fact that maximizing the remaining terms over $\theta$ will not affect the value of this constant, resulting in the argument maximizer over $\theta$ for these terms and for the entire likelihood $\mathcal{L}$ being identical. With this in mind, as well as the form of the $\theta$-containing

likelihood factor $L$ as derived at the end of the previous problem part, the process of deriving the MLE begins below by first log transforming. This is done due to the fact that this monotonic transform will not affect the argument maximizer, and the transformed likelihood is easier to work with.

$$L = \prod_{i=1}^{n} [\theta_1^{\delta_{1i}} e^{-\theta_1 x_{1i}}] \prod_{j=1}^{m} [\theta_2^{\delta_{2j}} e^{-\theta_2 x_{2j}}]$$
$$\implies \log(L) = \sum_{i=1}^{n} [\delta_{1i} \log(\theta_1) - \theta_1 x_{1i}] + \sum_{j=1}^{m} [\delta_{2j} \log(\theta_2) - \theta_2 x_{2j}]$$

This can now be maximized over $\theta = \{\theta_1, \theta_2\}$ using the derivative test as follows.

$$\begin{aligned}
\frac{\partial \log(L)}{\partial \theta_1} &= \frac{\partial}{\partial \theta_1} \{\sum_{i=1}^{n} [\delta_{1i} \log(\theta_1) - \theta_1 x_{1i}] + \sum_{j=1}^{m} [\delta_{2j} \log(\theta_2) - \theta_2 x_{2j}]\} \\
&= \sum_{i=1}^{n} [\delta_{1i} \frac{\partial}{\partial \theta_1} (\log(\theta_1)) - x_{1i} \frac{\partial}{\partial \theta_1} (\theta_1)] \\
&= \sum_{i=1}^{n} [\frac{\delta_{1i}}{\theta_1} - x_{1i}] \\
\frac{\partial \log(L)}{\partial \theta_2} &= \frac{\partial}{\partial \theta_2} \{\sum_{i=1}^{n} [\delta_{1i} \log(\theta_1) - \theta_1 x_{1i}] + \sum_{j=1}^{m} [\delta_{2j} \log(\theta_2) - \theta_2 x_{2j}]\} \\
&= \sum_{j=1}^{m} [\delta_{2j} \frac{\partial}{\partial \theta_2} (\log(\theta_2)) - x_{2j} \frac{\partial}{\partial \theta_2} (\theta_2)] \\
&= \sum_{j=1}^{m} [\frac{\delta_{2j}}{\theta_2} - x_{2j}]
\end{aligned}$$

Now, for the second derivatives as required to form the Hessian and check concavity, confirming the validity of the derivative test here.

$$\begin{aligned}
\frac{\partial^2 \log(L)}{\partial \theta_1^2} &= \frac{\partial}{\partial \theta_1} (\sum_{i=1}^{n} [\frac{\delta_{1i}}{\theta_1} - x_{1i}]) \\
&= \sum_{i=1}^{n} \delta_{1i} \frac{\partial}{\partial \theta_1} (\frac{1}{\theta_1}) \\
&= -\sum_{i=1}^{n} \frac{\delta_{1i}}{\theta_1^2}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \log(L)}{\partial \theta_2^2} &= \frac{\partial}{\partial \theta_2}(\sum_{j=1}^{m}[\frac{\delta_{2j}}{\theta_2} - x_{2j}]) \\
&= \sum_{j=1}^{m} \delta_{2j} \frac{\partial}{\partial \theta_2}(\frac{1}{\theta_2}) \\
&= -\sum_{j=1}^{m} \frac{\delta_{2j}}{\theta_2^2} \\
\frac{\partial^2 \log(L)}{\partial \theta_1 \partial \theta_2} &= \frac{\partial}{\partial \theta_2}(\sum_{i=1}^{n}[\frac{\delta_{1i}}{\theta_1} - x_{1i}]) \\
&= 0 \\
\frac{\partial^2 \log(L)}{\partial \theta_2 \partial \theta_1} &= \frac{\partial}{\partial \theta_1}(\sum_{j=1}^{m}[\frac{\delta_{2j}}{\theta_2} - x_{2j}]) \\
&= 0
\end{aligned}$$

The Hessian $H$ for $\log(L)$ is thus the following, using the definition of this matrix.

$$\begin{aligned}
H &= \begin{pmatrix} \frac{\partial^2 \log(L)}{\partial \theta_1^2} & \frac{\partial^2 \log(L)}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 \log(L)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \log(L)}{\partial \theta_2^2} \end{pmatrix} \\
&= \begin{pmatrix} -\sum_{i=1}^{n} \frac{\delta_{1i}}{\theta_1^2} & 0 \\ 0 & -\sum_{j=1}^{m} \frac{\delta_{2j}}{\theta_2^2} \end{pmatrix}
\end{aligned}$$

Note first that $\delta_{1i} \geq 0$ and $\delta_{2j} \geq 0$ for all $i, j$ by virtue of these being indicator functions. Next, recall that $\theta_1 > 0$ and $\theta_2 > 0$ due to these being exponential rate parameters. So, one finds that the Hessian of $\log(L)$ is a diagonal matrix with negative entries, which makes it inherently negative definite. As such, one finds that $\log(L)$ will be a strictly concave function, meaning that one can in fact find a maximum using the derivative test, setting the gradient of $\log(L)$ taken with respect to $\theta = \{\theta_1, \theta_2\}$ equal to zero and solving. This is undergone below to find the MLE $\hat{\theta} = \{\hat{\theta}_1, \hat{\theta}_2\}$ in terms of the observed data.

$$
\begin{aligned}
0 &= \frac{\partial \log(L)}{\partial \theta_1}|_{\theta=\hat{\theta}} \\
\implies 0 &= \sum_{i=1}^{n}[\frac{\delta_{1i}}{\hat{\theta}_1} - x_{1i}] \\
\implies \hat{\theta}_1 &= \frac{\sum_{i=1}^{n}\delta_{1i}}{\sum_{i=1}^{n}x_{1i}} \\
0 &= \frac{\partial \log(L)}{\partial \theta_2}|_{\theta=\hat{\theta}} \\
\implies 0 &= \sum_{j=1}^{m}[\frac{\delta_{2j}}{\hat{\theta}_2} - x_{2j}] \\
\implies \hat{\theta}_2 &= \frac{\sum_{j=1}^{m}\delta_{2j}}{\sum_{j=1}^{m}x_{2j}}
\end{aligned}
$$

So, the MLE for $\theta$ under each arm is $\boxed{\hat{\theta}_1 = \frac{\sum_{i=1}^{n}\delta_{1i}}{\sum_{i=1}^{n}x_{1i}}}$ for the control arm and $\boxed{\hat{\theta}_2 = \frac{\sum_{j=1}^{m}\delta_{2j}}{\sum_{j=1}^{m}x_{2j}}}$ for the treatment arm.

(c) The above formulae for $\hat{\theta}_1$ and $\hat{\theta}_2$ were implemented upon the full PBC data held within the survival package. First, the data was split into the two arms, removing all individuals who did not participate. Next, it was noted that, for this data, censoring was split into two categories, general and liver transplant. Both were considered indicators of censoring for the purposes of this analysis, such that $\delta = 0$ whenever either indicator takes non-zero value. A new variable was created to capture this. Finally, the MLEs were calculated using the expressions derived in the previous problem part. The code accomplishing this can be found below. Note that the tidyverse and survival libraries are imported for all code chunks, not just this first one.

```
library(tidyverse)
library(survival)

# Read data into environment
data("pbc")

# Collect estimates
treatment_data = pbc %>% filter(trt == 1)
placebo_data = pbc %>% filter(trt == 2)
```

```
# Create delta observations by grouping censoring causes
treatment_data = treatment_data %>%
  mutate(D = case_when(status == 2 ~ 1,
                       TRUE ~ 0))
placebo_data = placebo_data %>%
  mutate(D = case_when(status == 2 ~ 1,
                       TRUE ~ 0))

# Calculate actual parameter estimates
theta1_mle = sum(treatment_data$D)/sum(treatment_data$time)
theta2_mle = sum(placebo_data$D)/sum(placebo_data$time)
```

The resultant estimates for the hazard parameters can be found below.

$$\boxed{\hat{\theta}_1 = 0.0002041021}$$
$$\boxed{\hat{\theta}_2 = 0.0001951112}$$

To calculate the confidence intervals for these estimates, an asymptotic result was used. In particular, it was used that, as the amount of data grows, the Central Limit Theorem dictates that the distribution of $\hat{\theta}$ would approach a normal centered around $\theta$ with variance-covariance matrix equal to the inverse information $I(\theta)^{-1}$. Under regularity conditions, the information is the negative expectation of the Hessian of $\log(L)$. As this Hessian $H$ was calculated in the previous problem, $I(\theta)$ can be determined as follows. Note that $T_1, C_1$ are the general failure and censor time variables for the placebo arm, while $T_2, C_2$ are the same for the treatment arm.

$$\begin{aligned} I(\theta) &= E\begin{bmatrix} \sum_{i=1}^{n} \frac{\delta_{1i}}{\theta_1^2} & 0 \\ 0 & \sum_{j=1}^{m} \frac{\delta_{2j}}{\theta_2^2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{nP(T_1<C_1)}{\theta_1^2} & 0 \\ 0 & \frac{mP(T_2<C_2)}{\theta_2^2} \end{bmatrix} \end{aligned}$$

Plugging the best known estimates $\theta_1 \approx \hat{\theta}_1$, $P(T_1 < C_1) \approx \frac{1}{n}\sum_{i=1}^{n} \delta_{1i}$, $\theta_2 \approx \hat{\theta}_2$, and $P(T_2 < C_2) \approx \frac{1}{m}\sum_{i=1}^{m} \delta_{2j}$, one can estimate the information matrix to be the following.

$$I(\theta) \approx \begin{bmatrix} \frac{1}{\hat{\theta}_1^2}\sum_{i=1}^{n} \delta_{1i} & 0 \\ 0 & \frac{1}{\hat{\theta}_2^2}\sum_{j=1}^{m} \delta_{2j} \end{bmatrix}$$

Inverting this diagonal matrix can be accomplished by simply multiplicatively inverting the diagonal entries, resulting in the following variance-covariance matrix.

$$I(\theta)^{-1} \approx \begin{bmatrix} \frac{\hat{\theta}_1^2}{\sum_{i=1}^{n} \delta_{1i}} & 0 \\ 0 & \frac{\hat{\theta}_2^2}{\sum_{j=1}^{m} \delta_{2j}} \end{bmatrix}$$

As the variances of both $\hat{\theta}_1$ and $\hat{\theta}_2$ are their corresponding diagonal entries in this matrix, one finds from these asymptotic analyses that $\text{Var}(\hat{\theta}_1) \approx \frac{\hat{\theta}_1^2}{\sum_{i=1}^{n} \delta_{1i}}$ and $\text{Var}(\hat{\theta}_2) \approx \frac{\hat{\theta}_2^2}{\sum_{j=1}^{m} \delta_{2j}}$. Using these formulae for the approximate normal variance in the $\theta$ parameter estimates, a 95% confidence interval can be formed in the standard way. All of this is accomplished, along with the appropriate outputs, by the below code.

```
# Calculate variance of estimates using asymptotics
var_theta1 = theta1_mle^2/sum(treatment_data$D)
var_theta2 = theta2_mle^2/sum(placebo_data$D)

# Calculate and display confidence intervals
placebo = c(theta1_mle - 1.96*sqrt(var_theta1), theta1_mle,
    theta1_mle + 1.96*sqrt(var_theta1))
names(placebo) = c("Lower Bound", "Estimate", "Upper Bound")
placebo

treat = c(theta2_mle - 1.96*sqrt(var_theta2), theta2_mle,
    theta2_mle + 1.96*sqrt(var_theta2))
names(treat) = c("Lower Bound", "Estimate", "Upper Bound")
treat
```

This code produces the following confidence interval and estimate combined outputs. First, for the placebo group and $\hat{\theta}_1$.

```
 Lower Bound     Estimate  Upper Bound
0.0001544833 0.0002041021 0.0002537210
```

Finally, for the treatment group and the associated hazard parameter $\hat{\theta}_2$.

```
 Lower Bound     Estimate  Upper Bound
0.0001457412 0.0001951112 0.0002444811
```

(d) To plot the parametric survival function estimates for both arms, first recall the form of the survival function $S(t)$ for an $\text{Exp}(\theta)$-distributed failure time random, $S(t) = e^{-\theta t}$ variable. Using this general form, the parametric estimates for each arm, along with their 95% confidence intervals, could be plotted. For the estimates, one need only to plug in $\hat{\theta}_1$ and $\hat{\theta}_2$ into the

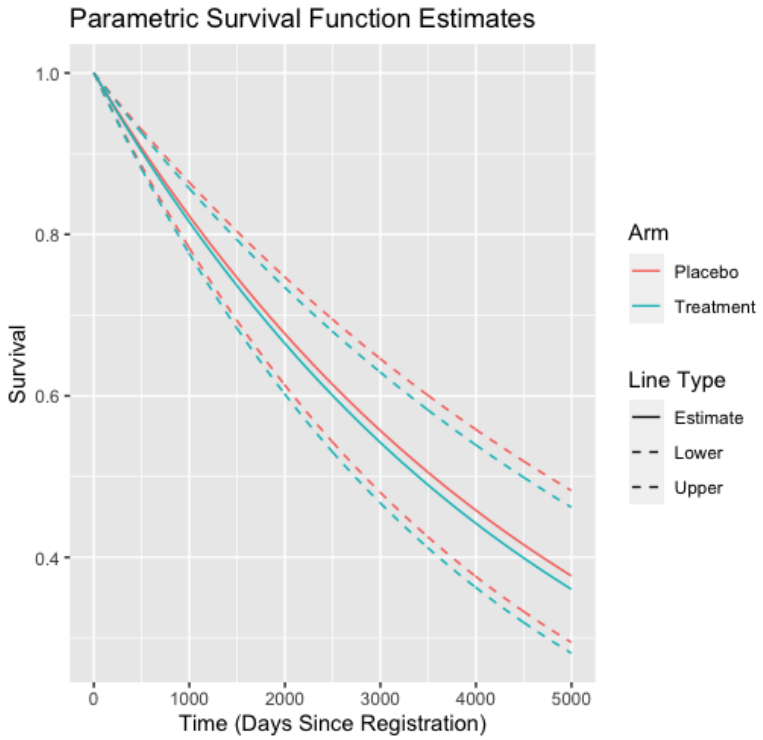
above form, with the upper and lower point-wise confidence bounds found by plugging in the corresponding bounds of the 95% confidence interval for each estimated parameter. This procedure is followed by the following code, where the resulting estimates for the two arms are overlaid on the same plot and differentiated by color.

```
# Create survival functions
survival_form <- function(x, theta){
  return(exp(-1*theta*x))
}

t = seq(0, 5000, by = 5)
plot_data1 = data.frame(time = t, Lower = survival_form(t, out1[1]),
                        Estimate = survival_form(t, out1[2]),
                        Upper = survival_form(t, out1[3]),
                        Arm = "Treatment") %>%
  pivot_longer(c(Lower, Estimate, Upper),
    names_to = "Line Type", values_to = "value")
plot_data2 = data.frame(time = t, Lower = survival_form(t, out2[1]),
                        Estimate = survival_form(t, out2[2]),
                        Upper = survival_form(t, out2[3]),
                        Arm = "Placebo") %>%
  pivot_longer(c(Lower, Estimate, Upper),
    names_to = "Line Type", values_to = "value")
plot_data = rbind(plot_data1, plot_data2)

# Plot survival functions with confidence intervals
ggplot(data = plot_data,
     aes(x = time, y = value, linetype = `Line Type`, color = Arm)) +
  geom_line() +
  scale_linetype_manual(values = c("solid", "dashed", "dashed")) +
  labs(x = "Time (Days Since Registration)", y = "Survival",
       title = "Parametric Survival Function Estimates")
```

Using the above code, the following plot was produced. Note that the confidence intervals are dashed lines rather than full.

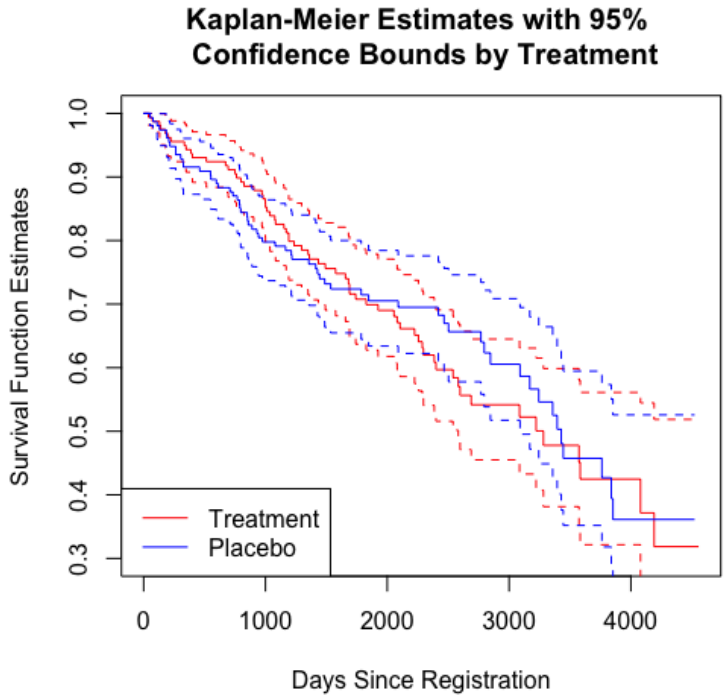



(e) The Kaplan-Meier curves color-coded by group, along with their corresponding point-wise confidence intervals in the same colors but with dotted lines, are plotted in the standard fashion as presented in the computing lab. This is accomplished using the following code, where the $\delta$ indicator is created for the PBC data, and all individuals within this dataset not within one of the study arms are exlcuded.

```
# Filter and create delta indicator
study_pop = pbc %>%
  filter(!is.na(trt)) %>%
  mutate(D = case_when(status == 2 ~ 1,
                       TRUE ~ 0))
# Create survival object
surv_obj = Surv(study_pop$time, study_pop$D, type = "right")
KM_by_treatment <- survfit(surv_obj ~ study_pop$trt)
KM_treat_summary <- summary(KM_by_treatment)
KM_treat_summary
plot(KM_by_treatment,
     conf.int = T,
     col = c("blue", "red"),
     main=("Kaplan-Meier Estimates with 95% \n
      Confidence Bounds by Treatment"),
```

```
      xlab = "Days Since Registration",
      ylab = "Survival Function Estimates",
      ylim = c(0.3, 1))
legend("bottomleft" ,legend = c("Treatment","Placebo"),
                     col = c("red", "blue"), lty = 1)
```

The above code block produces the following desired plot.




(f) The comparisons will be performed for each arm separately below.

Placebo/Control Arm:

The non-parametric KM estimate of the survival function has a less constrained descent compared to the exponential decay within the parametric estimate. In particular, it appears almost linear until around day 2000, after which there is very little drop in survival probability until the precipitous decline prior to day 3000. This type of plateau pattern, and other similar complex patterns can arise under the more flexible, non-parametric KM estimator, while it is impossible under the rigid parametric assumptions. Despite this difference though, the magnitudes of the two estimators, and their confidence intervals as well, are not all too different throughout the observed time period. The KM estimate is slightly above

the parametric one until somewhere between day 3000 and day 4000 of follow up.

As hinted at earlier, the main difference here is that there is simply a much greater deal of granularity and flexibility in the non-parametric KM estimate, which follows from its formulation and non-parametric nature. As a result of this, one can glean much more information from the KM estimate, such as the fact that there are relatively few uncensored survival times beyond 3000 days as indicated by the coarse nature of the graph past this point.

Treatment Arm:

For the treatment arm, one can make a very similar comparison. The treatment arm non-parametric KM estimate descends almost linearly, as opposed to the exponential descent enforced in the parametric model. However, despite this, again the magnitudes of the two estimators and their confidence intervals are not terribly different over time. The non-parametric estimate remains consistently slightly above the parametric one before approximately 2000 days, then dipping below after this point, a pattern one would expect from a negative-sloped linear and exponential decay functions both approximating the same process.

Again, one can see the greater level of data detail and flexibility in the KM estimate, with the same example holding here. Again one can observe that once again it can be observed that there are relatively few uncensored survival times beyond 3000 days as indicated by the coarse nature of the graph past this point.

(g) You can estimate the median survival time from Kaplain-Meier curves generally as the time $t$ where $\hat{S}(t) = 0.5$, such that estimated survival probability is $\frac{1}{2}$. However, this procedure requires that this value be attained by the survival function estimate $\hat{S}(t)$ within the observation period. As the KM curve estimates for both arms in fact attain this value, a fact evident in the plot in part (e), one can indeed estimate the median survival time using these Kaplain-Meier curves.

Using a similar procedure as above, one can estimate any quantile $q$ of the survival time where the relevant estimate $\hat{S}(t)$ achieves the value $1 - q$, such that the $q^{\text{th}}$ quantile can be found as $t$ such that $\hat{S}(t) = 1 - q$. This is the time where the estimated survival probability is $1 - q$, such that the probability of the survival time being less than or equal to $t$ is $q$, making this time the $q^{\text{th}}$ quantile as desired.

One other summary measure that provides information concerning the survival function estimate would be the area under the KM estimate curve. This value would estimate average survival time for the group whose survival function is estimated by the curve, which follows from the fact that

$E[T] = \int_0^\infty P(T > t)dt$ by probability results.

Another final piece of information one could extract would be the estimated linear slope of the KM survival estimates, however this type of information is somewhat reductive and requires an at least quasi-linear parametric assumption to provide any truly useful information.

# 2

(a) The log-rank test which will be performed has null Hypothesis $H_0 : \lambda_C = \lambda_T$ for hazards $\lambda_C$ of the control group and $\lambda_T$ of the treatment group. This will be a two-sided test, so the alternative hypothesis will be $H_1 : \lambda_C \neq \lambda_T$ for these same hazard functions, such that they are not the same function at any point. As is always the case for this test, it is assumed that the hazards do not crossover at any time $t$. This is somewhat problematic for this particular case, as crossover in survival functions implies crossover in hazards. As can be seen in part (e) of the previous problem, the estimates of the survival functions for the two arms indeed crossover multiple times. There is thus reason to believe that the non-crossover assumption on the hazards of these two groups is violated here, such that this test may not be entirely appropriate.

Moving forward despite the fact that it may not be entirely appropriate, the log-rank test can be performed for the PBC held in the survival library using the survdiff() function as indicated in the computing lab. First though, the data is filtered down to just those participants within one of the two study arms, and a censoring indicator is formed from the provided censoring data. The full procedure to perform the test is implemented through the following code block, and the resultant outputs follow.

```
# Filter and create delta indicator
study_pop = pbc %>%
  filter(!is.na(trt)) %>%
  mutate(D = case_when(status == 2 ~ 1,
                       TRUE ~ 0))

# Perform the relevant test
surv_diff <- survdiff(Surv(time, D, type = "right") ~ trt,
    data = study_pop)
surv_diff
```

The output produced by the above code can be found below.

```
Call:
survdiff(formula = Surv(time, D, type = "right") ~ trt, data = study_pop)

        N Observed Expected (O-E)^2/E (O-E)^2/V
trt=1 158       65     63.2    0.0502     0.102
trt=2 154       60     61.8    0.0513     0.102

 Chisq= 0.1  on 1 degrees of freedom, p= 0.7
```

As can be seen in the last line of the above output, performing a chi-squared Log-Rank test results in a statistic of 0.1 on 1 degree of freedom, which has p-value of 0.7. This test statistic is not very extreme for the $\chi^2$ distribution, resulting in the 0.7 p-value, which is quite far from the traditional significance threshold of p=0.05. As such, assuming that the hazards do not cross over, these results indicate that one cannot reject the null hypothesis that the hazards between these two groups are equal. This does seem to fit in with the nearly identical survival function estimates for the two arms, both parametric and Kaplan-Meier, presented in the previous problem. However, it must be taken with skepticism, as violation of the no crossover in hazards assumption, which is likely to occur here, results in the test statistic being systematically reduced toward zero, such that it will lose significance.

(b) First, note that fitting the desired Cox proportional hazards model will inherently exclude the data from the 106 individuals who did not participate in the randomized trial. This is due to their not being a part of the treatment nor the placebo arms. Despite not receiving the treatment, they in fact cannot be simply added in with the placebo arm participants, as the non-participants knew they were not receiving the treatment and never received any sort of placebo. Thus, mixing them with the placebo arm would both be dishonest and potentially introduce bias due to this prior knowledge of lack of treatment. Assigning these individuals to a third group, neither treatment nor placebo, would be acceptable, but then their data would not be used in the process of estimating the hazards ratio between the treatment and control arms anyway, so it is logical to just drop this data. Due to the data from these participants containing NA values in the trt field, the coxph() function call will actually drop these entries automatically.

Two models are fit here, one with just the treatment indicator and another including all given covariates to find the hazards ratio between the DPCA treatment and placebo controlling for the demographic and prognostic factors included in the dataset. To perform the these fits, the data for just the study participants must still be massaged slightly. This includes first creating an overall censoring variable, which indicates whether each participant was censored by any cause. Also, for the second model, the sex, Ascites, Hepatomegaly, Spiders, edema, and stage covariates must be converted to factors, as they take a finite number of discrete values.

Finally, for both models, the discrete treatment variable for all trial participants, which takes one of two values, was then also transformed into a factor rather than a continuous valued variable.

After the above data massaging was done, the coxph() function could be called to fit the two PHMs described above. Consider first the definition of the first model, with covariate $x$ containing an indicator of treatment in the following fashion.

$$x = \begin{cases} 1 \text{ if in treatment arm} \\ 0 \text{ if in placebo arm} \end{cases}$$

Using this binary covariate, a parameter $-\infty < \beta < \infty$, and a non-parametric baseline hazard function $\lambda_0(t)$, the first PHM described above can be formulated as follows.

$$\lambda(t; x) = \lambda_0(t)e^{\beta x}$$

For the second model, all that changes is the covariate and parameter structure. For the first change, consider $17 \times 1$ covariate vector $z$ defined in the following fashion.

$$z_1 = \begin{cases} 1 \text{ if in treatment arm} \\ 0 \text{ if in placebo arm} \end{cases}$$
$$z_2 = \text{age of subject}$$
$$z_3 = \text{sex of subject}$$
$$z_4 = \text{Ascites Presence}$$
$$z_5 = \text{Hepatomegaly Presence}$$
$$z_6 = \text{Spiders Presence}$$
$$z_7 = \text{Edema Presence}$$
$$z_8 = \text{Bilirubin}$$
$$z_9 = \text{Cholesterol}$$
$$z_{10} = \text{Albumin}$$
$$z_{11} = \text{Urine Copper}$$
$$z_{12} = \text{Alkaline Phosphatase}$$
$$z_{13} = \text{SGOT}$$
$$z_{14} = \text{Triglycerides}$$
$$z_{15} = \text{Platelet Count}$$
$$z_{16} = \text{Prothrombin Time}$$
$$z_{17} = \text{Histological Stage}$$

Using this covariate vector, an unconstrained parameter vector $\gamma$ which is $17 \times 1$, and another non-parametric baseline hazard function $\eta_0(t)$, the second PHM described above, which will be used to find the hazards ratio between the DPCA treatment and placebo controlling for all other collected covariates, can be expressed as follows, where $\eta(t; x)$ is the hazard under this model at time $t$ for an individual having covariate $x$.

$$\eta(t; x) = \eta_0(t) e^{\gamma' z}$$

The code to fit these two models, and output the resultant fits, can be found below.

```
# Set up the data for fit
full_data = pbc %>%
  mutate(D = case_when(status == 2 ~ 1,
                       TRUE ~ 0)) %>%
  filter(!is.na(trt)) %>%
  mutate(sex = as.factor(sex),
         ascites = as.factor(ascites),
         hepato = as.factor(hepato),
```

```
            spiders = as.factor(spiders),
            edema = as.factor(edema),
            stage = as.factor(stage),
            trt = as.factor(trt))

# Perform the first model fit
fit1 = coxph(Surv(time, D, type = "right") ~ trt,
             control = coxph.control(iter.max = 50),
             ties= "breslow", data = full_data)

# Perform the second model fit, controlling for covariates
fit2 = coxph(Surv(time, D, type = "right") ~ trt + age + sex + ascites +
              hepato + spiders + edema + bili + chol + albumin + copper +
              alk.phos + ast + trig + platelet + protime + stage,
            control = coxph.control(iter.max = 50),
            ties= "breslow", data = full_data)

# Analyze the fit results
summary(fit1)
summary(fit2)
```

The summary of the fit for the first model mentioned above, with just the treatment covariate in the model, can be found below.

```
Call:
coxph(formula = Surv(time, D, type = "right") ~ trt, data = full_data,
    control = coxph.control(iter.max = 50), ties = "breslow")

  n= 312, number of events= 125

        coef exp(coef) se(coef)      z Pr(>|z|)
trt2 -0.05712   0.94448  0.17917 -0.319     0.75

     exp(coef) exp(-coef) lower .95 upper .95
trt2    0.9445      1.059    0.6648     1.342

Concordance= 0.499  (se = 0.025 )
Likelihood ratio test= 0.1  on 1 df,   p=0.7
Wald test            = 0.1  on 1 df,   p=0.7
Score (logrank) test = 0.1  on 1 df,   p=0.7
```

Next, the summary of the fit for the second model mentioned above, controlling for all other present covariates, can be found below.

```
Call:
coxph(formula = Surv(time, D, type = "right") ~ trt + age + sex +
    ascites + hepato + spiders + edema + bili + chol + albumin +
    copper + alk.phos + ast + trig + platelet + protime + stage,
    data = full_data, control = coxph.control(iter.max = 50),
    ties = "breslow")

  n= 276, number of events= 111
   (36 observations deleted due to missingness)

               coef  exp(coef)   se(coef)      z Pr(>|z|)
trt2     -1.762e-01  8.384e-01  2.189e-01 -0.805  0.42073
age       2.896e-02  1.029e+00  1.164e-02  2.488  0.01285 *
sexf     -3.756e-01  6.869e-01  3.111e-01 -1.207  0.22733
ascites1 -2.953e-04  9.997e-01  3.951e-01 -0.001  0.99940
hepato1   5.699e-02  1.059e+00  2.565e-01  0.222  0.82420
spiders1  7.034e-02  1.073e+00  2.491e-01  0.282  0.77770
edema0.5  2.422e-01  1.274e+00  3.355e-01  0.722  0.47045
edema1    1.146e+00  3.146e+00  4.167e-01  2.751  0.00594 **
bili      8.013e-02  1.083e+00  2.629e-02  3.048  0.00230 **
chol      4.737e-04  1.000e+00  4.534e-04  1.045  0.29614
albumin  -7.479e-01  4.733e-01  3.103e-01 -2.410  0.01595 *
copper    2.445e-03  1.002e+00  1.170e-03  2.089  0.03670 *
alk.phos  6.162e-07  1.000e+00  4.066e-05  0.015  0.98791
ast       3.706e-03  1.004e+00  1.992e-03  1.860  0.06283 .
trig     -5.327e-04  9.995e-01  1.436e-03 -0.371  0.71069
platelet  8.415e-04  1.001e+00  1.189e-03  0.708  0.47902
protime   2.765e-01  1.319e+00  1.168e-01  2.368  0.01786 *
stage2    1.408e+00  4.086e+00  1.080e+00  1.303  0.19255
stage3    1.683e+00  5.384e+00  1.052e+00  1.600  0.10954
stage4    2.119e+00  8.327e+00  1.068e+00  1.985  0.04711 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
          exp(coef) exp(-coef) lower .95 upper .95
trt2         0.8384     1.1927    0.5460    1.2876
age          1.0294     0.9715    1.0062    1.0531
sexf         0.6869     1.4558    0.3733    1.2638
ascites1     0.9997     1.0003    0.4609    2.1685
hepato1      1.0586     0.9446    0.6403    1.7503
spiders1     1.0729     0.9321    0.6584    1.7483
edema0.5     1.2740     0.7849    0.6601    2.4590
edema1       3.1464     0.3178    1.3904    7.1203
bili         1.0834     0.9230    1.0290    1.1407
chol         1.0005     0.9995    0.9996    1.0014
albumin      0.4733     2.1126    0.2576    0.8696
copper       1.0024     0.9976    1.0002    1.0048
alk.phos     1.0000     1.0000    0.9999    1.0001
ast          1.0037     0.9963    0.9998    1.0076
trig         0.9995     1.0005    0.9967    1.0023
platelet     1.0008     0.9992    0.9985    1.0032
protime      1.3186     0.7584    1.0488    1.6576
stage2       4.0861     0.2447    0.4919   33.9449
stage3       5.3842     0.1857    0.6850   42.3231
stage4       8.3267     0.1201    1.0274   67.4827

Concordance= 0.848  (se = 0.018 )
Likelihood ratio test= 169.6  on 20 df,   p=<2e-16
Wald test            = 174.9  on 20 df,   p=<2e-16
Score (logrank) test = 295.3  on 20 df,   p=<2e-16
```

To calculate the hazard ratio between the treatment and placebo groups for either of the models above, one must calculate the hazard ratio between two individuals, identical across all covariates aside from one being treated and the other placebo. This ratio simplifies under the first model to be $\frac{\lambda(t;x=1)}{\lambda(t;x=0)} = e^{\beta_1}$ and under the second model to be $\frac{\eta(t;z_1=1)}{\eta(t;z_1=0)} = e^{\gamma_1}$ using the notation above. So, all that must be done is find the exponential of the covariates $\beta_1$ and $\gamma_1$ for the treatment indicator in each model above. As treatment is indicated by trt = 2 in this data, the first model has $\beta_1 = -0.05712$ and $e^{\beta_1} = 0.9445$, with the 95% confidence interval for this hazard ratio being (0.6648, 1.342). Using the same approach, one can find the second model has $\gamma_1 = -0.1762$ with $e^{\gamma_1} = 0.8384$ and confidence interval of the hazard ratio being (0.5460, 1.2876). Under both models, leaving out the other covariates and accounting for them, the 95% confidence interval for the hazard ratio between the treatment and placebo groups most certainly contains 1, it is in fact quite close to the center of both intervals, and the hazard ratios themselves are quite close to

1. As a result of the 95% confidence interval containing 1, one cannot say with the traditional 0.05 certainty level that the DPCA treatment reduces hazard/improves survival under either model.

(c) First, note that some of the given demographic and prognostic factors which were not in the list of basic measures, in particular Alkaline Phosphatase, Ascites, Hepatomegaly, and Spiders. As such, all of these data are missing for the individuals who did not participate in the clinical trial. As such, the data from these individuals will have to be excluded for this analysis as well. After performing this exclusion, the data was again massaged in the same fashion as done for the previous problem part to facilitate ease of PHM fit. After doing this massaging, a single Cox PHM was fit to the resultant data, with the only covariates being those that are requested. This will provide the effect of each of the desired demographic and prognostic covariates upon hazard, and thus survival, controlling for the other desired covariates. This does mean that several covariates, including treatment status and stage of disease, will not be controlled for in this model. This is simply the interpretation taken for the question asked, as the associations given treatment or given stage of disease is not requested, just the associations themselves. With all of this in mind, the requisite data massaging, model fitting, and observation of resulting model fit results are undergone using the below code.

```
# Set up the data for fit
full_data = pbc %>%
  mutate(D = case_when(status == 2 ~ 1,
                       TRUE ~ 0)) %>%
  filter(!is.na(trt)) %>%
  mutate(sex = as.factor(sex),
         ascites = as.factor(ascites),
         hepato = as.factor(hepato),
         spiders = as.factor(spiders),
         edema = as.factor(edema),
         stage = as.factor(stage),
         trt = as.factor(trt))

# Perform the fit
fit = coxph(Surv(time, D, type = "right") ~ age + sex + edema + bili +
              albumin + platelet + protime + alk.phos + ascites + hepato
              + spiders ,
            control = coxph.control(iter.max = 50),
            ties= "breslow", data = full_data)

# Analyze the fit results
summary(fit)
```

The resulting regression coefficients (coef), standard errors (se(coef)), and

p-values (P(> |z|)) for the model fit above can be found below.

```
Call:
coxph(formula = Surv(time, D, type = "right") ~ age + sex + edema +
    bili + albumin + platelet + protime + alk.phos + ascites +
    hepato + spiders, data = full_data %>% filter(trt != 3),
    control = coxph.control(iter.max = 50), ties = "breslow")

  n= 308, number of events= 124
   (4 observations deleted due to missingness)

              coef  exp(coef)   se(coef)      z Pr(>|z|)
age       2.891e-02  1.029e+00  9.737e-03  2.969 0.002988 **
sexf     -4.485e-01  6.386e-01  2.681e-01 -1.673 0.094384 .
edema0.5 -6.150e-03  9.939e-01  2.900e-01 -0.021 0.983081
edema1    8.548e-01  2.351e+00  3.296e-01  2.594 0.009498 **
bili      1.094e-01  1.116e+00  1.697e-02  6.447 1.14e-10 ***
albumin  -9.686e-01  3.796e-01  2.674e-01 -3.621 0.000293 ***
platelet -6.250e-04  9.994e-01  1.021e-03 -0.612 0.540388
protime   2.369e-01  1.267e+00  8.556e-02  2.769 0.005623 **
alk.phos  1.992e-05  1.000e+00  3.549e-05  0.561 0.574548
ascites1  2.491e-01  1.283e+00  3.113e-01  0.800 0.423568
hepato1   5.026e-01  1.653e+00  2.199e-01  2.285 0.022294 *
spiders1  2.516e-01  1.286e+00  2.128e-01  1.183 0.237001
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As can be seen above, the Age, full presence of Edema, Bilirubin, Albumin, Prothrombin Time, and presence of Hepatomegaly coefficients achieve the level of p=0.05 statistical significance. Of these covariates which have a statistically significant values, age, Edema, Bilirubin, Prothrombin Time, and Hepatomegaly all have hazard ratios greater than 1 for a positive 1 unit change in their value, controlling for other relevant measures. As such, these covariates have a statistically significant increase in hazard, and reduction in survival, as they are increased in value. For the binary variates, this means that there is an increase in hazard, and reduction in survival, with the presence of Edema as well as the presence of Hepatomegaly over the base conditions of each of these not being present.

On the other hand, Albumin has a hazard ratio less than 1 for a positive 1 unit change in its value, controlling fore the other relevant measures.

(d) This treatment of liver transplantation as censoring is not appropriate. Only those individuals for whom PBC has progressed to a great degree, such that it becomes life threatening, will be considered for such transplants, such that there will be inherent correlation and dependence between the censoring time due to liver transplant and the failure time of death. This will consistently be the case even when choosing individuals at the same coviarate values. These values do not paint a full picture of the health of the participants, so the same correlation and dependence

described above will be present. As such, the assumption of independent censoring given the covariates, which is required for validity of the PHM and other analyses performed in previous problem parts, is violated when liver transplantation is treated as censoring. This means the above analyses should be treated with skepticism.