Survival Analysis - Winter 2024
Assignment 2

1.

(a) From the description, we know that $S_0(t)$ is the survival function of the random variable $\exp(\alpha + \sigma\epsilon)$, so we have:

$$\begin{aligned} S_0(t) &= P(\exp(\alpha + \sigma\epsilon) > t) \\ &= P(\exp(\alpha + \sigma\epsilon)\exp(\gamma x_i)\exp(-\gamma x_i) > t) \\ &= P(\exp(\alpha + \sigma\epsilon + \gamma x_i) > t\exp(\gamma x_i)) \quad \text{[note: rearranging terms]} \\ &= S_i(t\exp(\gamma x_i)) \quad \text{[note: } S_i(t) \text{ is the survival function of RV } \exp(\alpha + \sigma\epsilon + \gamma x_i)] \end{aligned}$$

(b) Similarly,

$$\begin{aligned} S_i(t) &= P(\exp(\alpha + \sigma\epsilon + \gamma x_i) > t) \quad \text{[note: by definition]} \\ &= P(\exp(\alpha + \sigma\epsilon) > t\exp(-\gamma x_i)) \\ &= S_0(t\exp(-\gamma x_i))) \end{aligned}$$

The hazard function can be derived from the survival function as $\lambda(t) = \frac{-lnS(t)}{dt}$, Given,

$$S_i(t) = S_0(t\exp(-\gamma x_i))$$

We take the derivatives w.r.t t on both sides of the equation, which gives the following:

$$\begin{aligned} -\frac{dln(S_i(t))}{dt} &= -\frac{dlnS_0(t\exp(-\gamma x_i))}{dt} \\ &= -\frac{dlnS_0(t\exp(-\gamma x_i))}{d(t\exp(-\gamma x_i))}\frac{d(t\exp(-\gamma x_i))}{dt} \\ &= \lambda_0(t\exp(-\gamma x_i))\exp(-\gamma x_i) \quad [-\frac{dlnS_0(t\exp(-\gamma x_i))}{d(t\exp(-\gamma x_i))} \text{ by defn is hazard } \lambda_0(.)] \end{aligned}$$

and by definition, the left hand side $-\frac{dln(S_i(t))}{dt}$ is $\lambda_i(t)$, therefore we showed that:

$$\lambda_i(t) = \lambda_0(t\exp(-\gamma x_i))\exp(-\gamma x_i)$$

(c) The relationship between survival functions $S_i(t)$ and $S_\epsilon(t)$ can be derived as follows:

$$\begin{aligned} S_i(t) &= P(\exp(\alpha + \sigma\epsilon + \gamma x_i) > t) \quad \text{[note: by definition]} \\ &= P(\exp(\sigma\epsilon) > t\exp(-\alpha - \gamma x_i)) \\ &= P(\sigma\epsilon > \ln(t) - \alpha - \gamma x_i) \quad \text{[note: take the log on both sides]} \\ &= S_\epsilon(\frac{\ln(t) - \alpha - \gamma x_i}{\sigma}) \end{aligned}$$

Which gives:

$$g(t) = \frac{\ln(t) - \alpha - \gamma x_i}{\sigma}$$

where $\sigma \neq 0$.

2.

(a) The Kaplan-Meier estimator is given by

$$\begin{aligned}
\hat{S}_{KM}(t) &= \prod_{j:t(j)\leq t} (1 - \frac{d_j}{n_j}) \\
&= \prod_{j:t(j)\leq t} (\frac{n_j - d_j}{n_j}) \\
&= (\frac{n - d_1}{n})(\frac{n - d_1 - d_2}{n - d_1})(\frac{n - d_1 - d_2 - d_3}{n - d_1 - d_2})...(\frac{n - d_1 - d_2 - .... - d_j}{n - d_1 - d_2 - ....d_{j-1}}) \\
&= \frac{n - d_1 - d_2 - .... - d_j}{n} \\
&= 1 - \frac{\sum_{j:t(j)\leq t} d_j}{n} \quad \text{[note: } \sum_{j:t(j)\leq t} d_j \text{ is the total number of events for t(j)} \leq \text{t]} \\
&= 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{t_i < t\} \quad \text{[note:} \sum_{i=1}^{n} \mathbf{1}\{t_i < t\} = \sum_{j:t(j)\leq t} d_j] \\
&= 1 - \hat{F}(t)
\end{aligned}$$

(b)

$$\begin{aligned}
\hat{Var}(\hat{S}_{KM}(t)) &= Var(1 - \hat{F}(t)) \quad \text{note:[we've shown that } S_{KM}(t) = 1 - \hat{F}(t)] \\
&= Var(\hat{F}(t)) \\
&= Var(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{t_i \leq t\}}) \quad \text{[note: } \hat{F}(t) \text{ is the empirical CDF]} \\
&= \frac{1}{n^2} Var(\sum_{i=1}^{n} \mathbf{1}_{\{t_i \leq t\}}) \text{ [} \mathbf{1}_{\{t_i \leq t\}} \sim Bern(P(t_i \leq t)), \text{by defn,} P(t_i \leq t) = \hat{F}(t)] \\
&= \frac{1}{n^2} n\hat{F}(t)(1 - \hat{F}(t))[\sum_{i=1}^{n} \mathbf{1}_{\{t_i \leq t\}} \text{ sum of Bernulli,} \sum_{i=1}^{n} \mathbf{1}_{\{t_i \leq t\}} \sim Bin(n, \hat{F}(t)) \text{]} \\
&= \frac{\hat{F}(t)(1 - \hat{F}(t))}{n} \\
&= \frac{(1 - \hat{S}_{KM}(t))\hat{S}_{KM}(t)}{n}
\end{aligned}$$

Or alternatively, we can show $\frac{\hat{V}(S_{KM}(t))}{\hat{S}_{KM}(t)^2}$ and $\frac{1-\hat{S}_KM(t)}{n\hat{S}_{KM}(t)}$are equal.

$$\begin{aligned}
\frac{\hat{V}(S_{KM}(t))}{\hat{S}_{KM}(t)^2} &= \sum_{j:t_j\le t}\frac{d_j}{n_j(n_j-d_j)} \quad \text{[note: by Greeenwood formula]} \\
&= \sum_{j:t_j\le t}\frac{1}{(n_j-d_j)}-\frac{1}{n_j} \\
&= (\frac{1}{n-d_1}-\frac{1}{n})+(\frac{1}{(n-d_1)-d_2}-\frac{1}{n-d_1}) \\
&+(\frac{1}{(n-d_1-d_2)-d_3}-\frac{1}{n-d_1-d_2})... \\
&+(\frac{1}{n-\sum_{j:t_j\le t}d_j}-\frac{1}{n-\sum_{j-1:t_j-1\le t}d_j}) \\
&= \frac{1}{n-\sum_{j:t_j\le t}d_j}-\frac{1}{n} \\
&= \frac{1}{n(1-\hat{F}(t))}-\frac{1}{n} \quad \text{[note:}\hat{F}(t)=\sum_{i=1}^{n}\mathbf{1}\{t_i<t\}=\sum_{j:t(j)\le t}d_j\text{]}
\end{aligned}$$

$$\begin{aligned}
\frac{1-\hat{S}_KM(t)}{n\hat{S}_{KM}(t)} &= \frac{1}{n\hat{S}_KM(t)}-\frac{1}{n} \\
&= \frac{1}{n(1-\hat{F}(t))}-\frac{1}{n}
\end{aligned}$$

We showed that

$$\frac{\hat{V}(S_{KM}(t))}{\hat{S}_{KM}(t)^2}=\frac{1-\hat{S}_KM(t)}{n\hat{S}_{KM}(t)}=\frac{1}{n(1-\hat{F}(t))}-\frac{1}{n}$$

and we also know:

$$\hat{S}_{KM}(t)=1-\hat{F}(t)$$

which gives us:

$$\frac{S_{KM}(t)(1-S_{KM}(t))}{n}=\frac{\hat{F}(t)(1-\hat{F}(t))}{n}$$

So we showed that

$$V(S_{KM}(t)=\frac{S_{KM}(t)(1-S_{KM}(t))}{n}=\frac{\hat{F}(t)(1-\hat{F}(t))}{n}$$

3.

(a)

$$g(S_{KM}(t)) = log(-log(S_{KM}(t))$$

take derivative w.r.t $S_{KM}(t)$

$$\begin{aligned} g'(S_{KM}(t)) &= (\frac{1}{log(S_{KM}(t)^{-1})})(\frac{1}{S_{KM}(t)^{-1}})(-S_{KM}(t)^{-2}) \\ &= \frac{1}{S_{KM}(t)log(S_{KM}(t))} \end{aligned}$$

by Delta method $V[g(\hat{\theta})] \approx (g'(\theta))^2 V[\hat{\theta}]$. Therefore,

$$\begin{aligned} V[g(\hat{S}_{KM}(t))] &= (g'(S_{KM}(t)))^2 V(\hat{S}_{KM}(t)) \\ &= (\frac{1}{S_{KM}(t)log(S_{KM}(t))})^2(\frac{\hat{S}_{KM}(t)^2(1-\hat{S}_{KM}(t))}{n}) \end{aligned}$$

Then the 95% CI can be derived as follows:

$$-1.96 \leq \frac{g(\hat{S}_{KM}(t)) - g(S_{KM}(t))}{g'(\hat{S}_{KM}(t))\sqrt{V(\hat{S_{KM}}(t))}} \leq 1.96$$

$$-1.96 \leq \frac{log(-log(S_{KM}(t)) - log(-log(\hat{S}_{KM}(t))}{g'(\hat{S}_{KM}(t))\sqrt{V(\hat{S_{KM}}(t))}} \leq 1.96$$

Let

$$se = g'(\hat{S}_{KM}(t))\sqrt{V(\hat{S_{KM}}(t))}$$

So the 95% CI for $log(-log(S_{KM}(t))$ is:

$log(-log(\hat{S}_{KM}(t))-1.96se \leq log(-log(S_{KM}(t)) \leq log(-log(\hat{S}_{KM}(t))+1.96se$

Let

$$L = log(-log(\hat{S}_{KM}(t)) - 1.96se$$

and

$$U = log(-log(\hat{S}_{KM}(t)) + 1.96se$$

therefore we have:

$$L \leq log(-log(S_{KM}(t)) \leq U$$

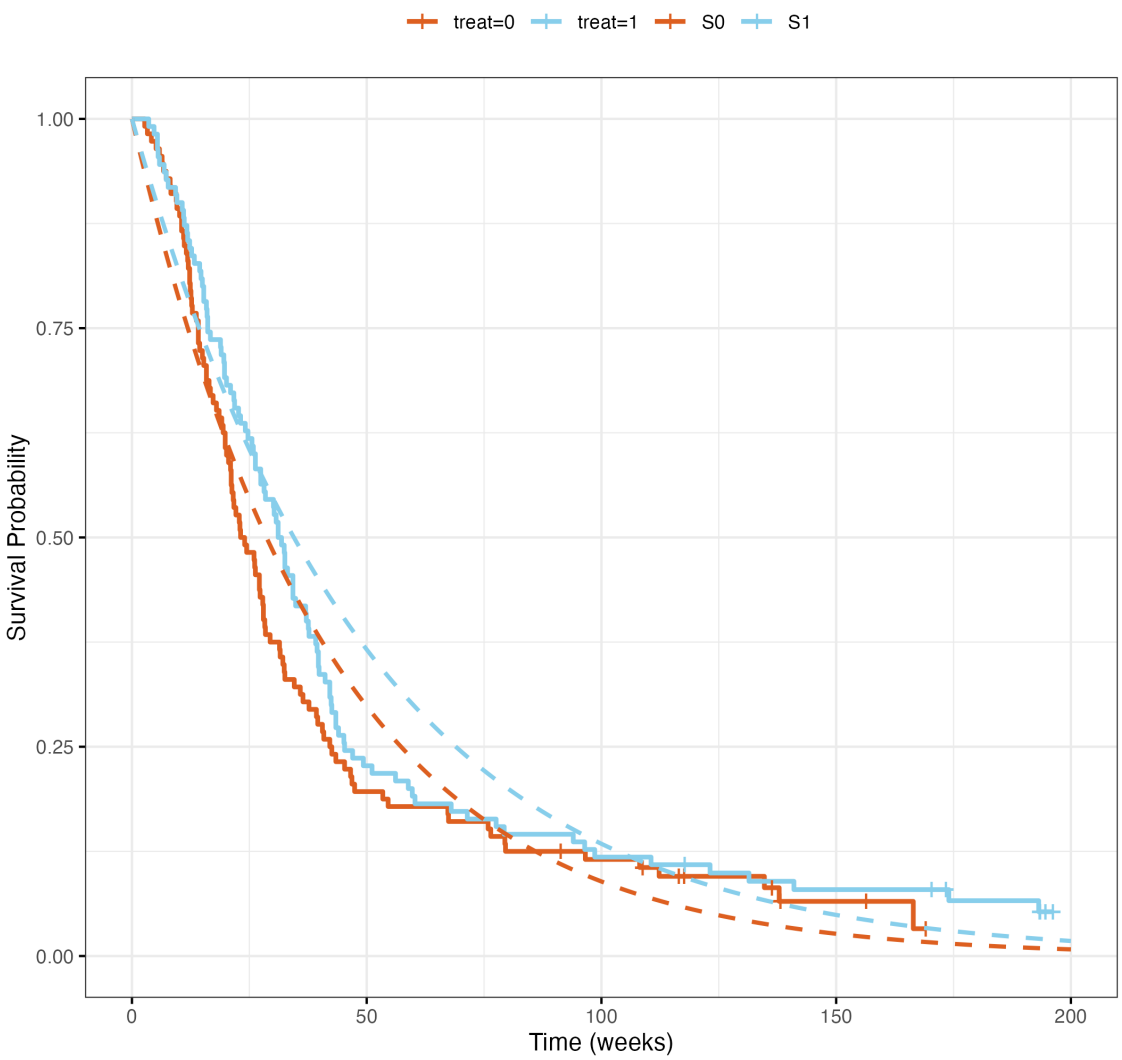
Then 95% CI for $S_{KM}(t)$ can be written as:

$$\exp(-\exp(U)) \leq S_{KM}(t) \leq \exp(-\exp(L))$$

we know that $f(x) = \exp(x)$ takes the values in (0,∞), so $-\exp(x)$ takes the values in (-∞,0), it then gives that $\exp(-\exp(x))$ takes the values between 0 and 1. Therefore, the CI of $S_{KM}(t)$ is bounded by [0,1] using this log log transformation.

4.

(a) The Kaplan-Meier curves (solid lines) and survival curves from exponential model (dashed lines) are presented below (codes are attached in the appendix). The blue lines are treat=1(chemo) groups and orange lines are treat=0(placebo) group:



We can see from the plot that:

- general speaking, the exponential model fit the data well as we see the estimated survival probabilities from the exponential model are close to/overlap with the Kaplan-Meier curves in some regions and the trends/curvatures are similar. But on the other hand, the KM curves shows that the survival between treat=0(placebo) and treat=1(chemo) groups are very similar, with treat=1(chemo) group has slightly better survival, however, the exponential model shows

that treat=1(chemo) group always has better survival comparing to treat=0(placebo) group.

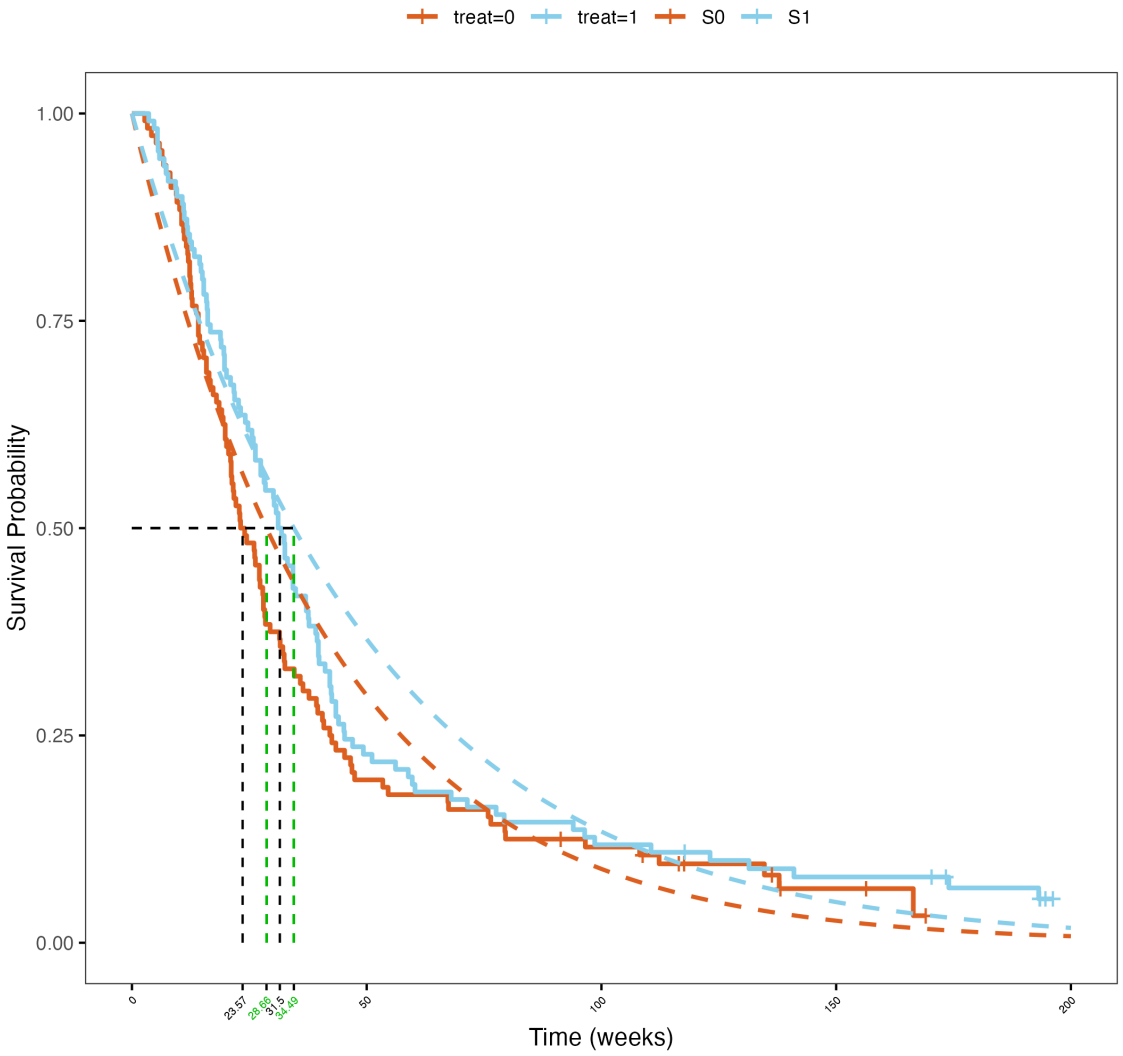
- the exponential model tends to overestimate the survival probabilities between the 50^{th} week and the 100^{th} (roughly) week, and underestimate the survival probabilities from the $100^{th}/125^{th}$ week to the end, but the difference is not much.
- And we can see that the exponential curves fit the KM curves better before the median survival time(s). and the difference between exponential model and KM estimation gets larger after the median survival time(s).

(b) median survival time can be found at where the survival probability is 50% . For the exponential model, we set s=0.5, and use the `uniroot` function to derive the median survival time. Using the following codes, we are able to get the median survival times:

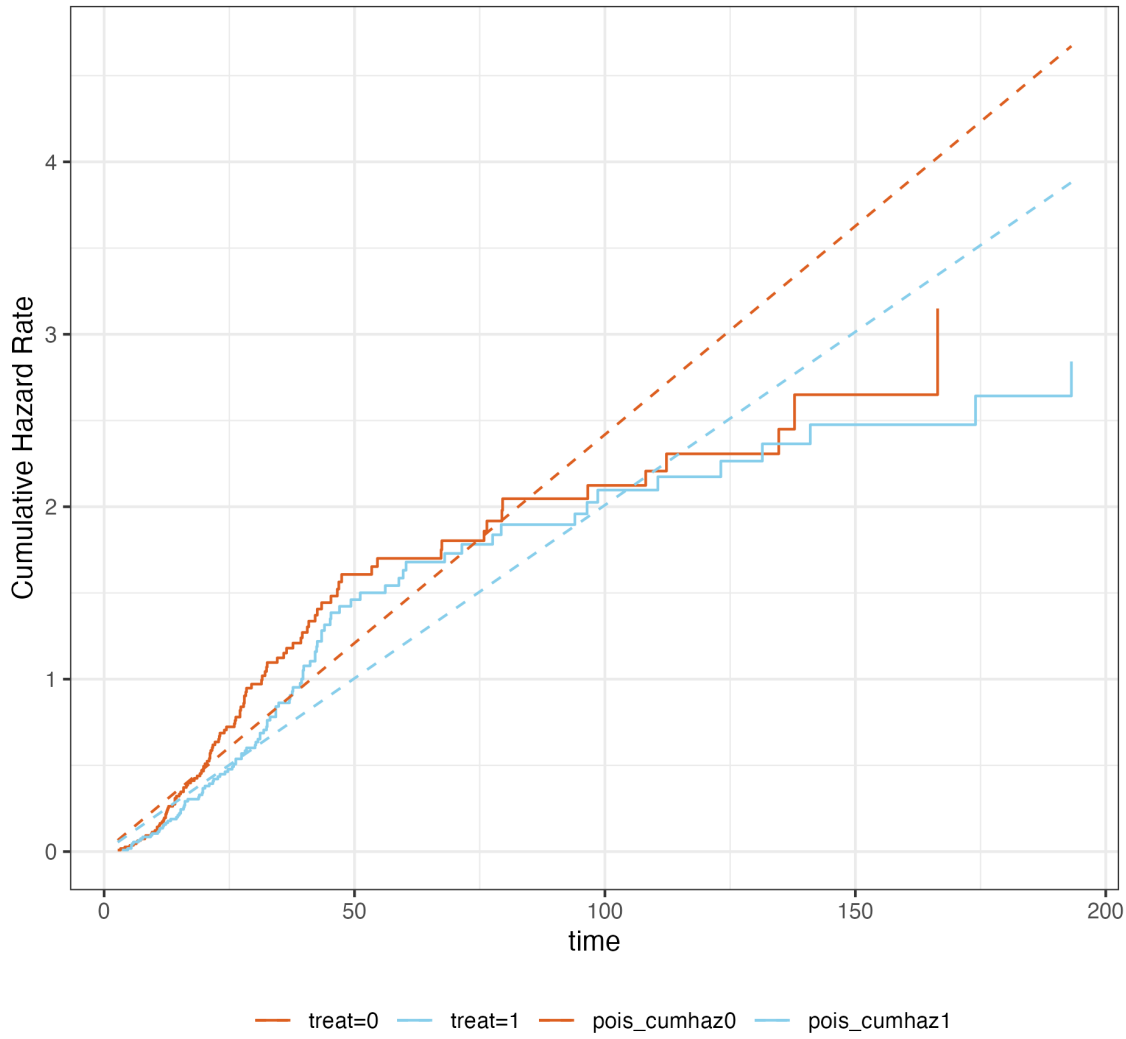
```
## median survival time exponential model
S0 <- function(t,s){exp(-(t*lambda0))-s}
S1 <- function(t,s){exp(-(t*lambda0*exp(beta)))-s}
t0 <- uniroot(S0, s=0.5, interval=c(0,200))$root %>% round(1)
t1 <- uniroot(S1, s=0.5, interval=c(0,200))$root %>% round(1)

## median survival time KM curve
km_median_dat <- summary(fit_treat)$table %>% data.frame()
km_t0 <- km_median_dat$median[1]
km_t1 <- km_median_dat$median[2]
```

The following plot shows the median survival times (codes generating the plot can be found in the appendix). The green dashed vertical lines and numbers correspond to the median survival times from the exponential models, and the black dashed vertical lines and numbers are from the Kaplan-Meier estimation. The median survival times are summarized as bellow:

	KM	exponential	diff
treat=0(placebo)	23.57	28.66	5.09
treat=1(chemo)	31.50	34.49	2.99



(c) The cumulative hazard rates are shown in the following plot (codes can be found in the appendix). the solid lines are Nelson-Aalen estimates, and the dashed lines are the estimates from exponential model. the orange curves are for treat=0(placebo) group and blue curves for treat=1(chemo) group. We can see the exponential estimates are different from the Nelson-Aelen estimates. The Nelson-Aalen curves showed that the hazard is not constant over time(the slope decreases over time), while the exponential model assumes constant hazard. So the model does not fit the data well as the assumption of constant hazard is not met based on the visual.



5.

(a) • The Weibull model can be specified as $\lambda_i(t;\theta) = \lambda\kappa(\lambda t)^{\kappa-1}\exp(\beta \text{treat})$:

```
> model_wei_ph <- phreg(Surv(weeks, event) ~ treat,
+                       dist = "weibull", data=brain)
> model_wei_ph_dat <- summary(model_wei_ph) %>%
+   coefficients() %>% data.frame()
>
> lower <- exp(model_wei_ph_dat$coef -1.96*model_wei_ph_dat$se.coef.)
> upper <- exp(model_wei_ph_dat$coef +1.96*model_wei_ph_dat$se.coef.)
>
> ## hazard ratio
> model_wei_ph_dat$exp.coef. %>% round(2)
```

```
[1] 0.82
## lower CI
> lower %>% round(2)
[1] 0.63
> ## upper CI
> upper %>% round(2)
[1] 1.08
```

- The hazard ratio can be computed as $exp(\beta) = 0.82$, and 95% CI, which can be computed as $\exp(\beta + se(\beta))$, is (0.63,1.08).
- The results can be interpreted as: The hazard rate in the treatment group (chemo) is 0.82 times that in the placebo group. So the chemo treatment reduced the risk of death by 18%, however the 95% confidence interval for the rate ratio is(0.63,1.08),which cross 1, it indicates that the treatment effect is not statistically significant (at an α level of 0.05)
- The exponential and Weibull models are fitted, LR test is conducted to compare the two models. The LR statistics follows a chi-squared distribution with df=1, LR = 0.76 and the corresponding P-value is 0.38, which indicates that the exponential and Weibull model are not statistically different. So we conclude that the exponential model fit the data adequately.

```
## exponential model
model_aft_exp <- survreg(Surv(weeks, event) ~ treat,
                     dist="exponential", data=brain)
summary(model_aft_exp)

## weibull model
model_aft_wei <- survreg(Surv(weeks, event) ~ treat,
                         dist="weibull", data=brain)
summary(model_aft_exp)

> -2*(model_aft_exp$loglik[2]-model_aft_wei$loglik[2])
[1] 0.7582267
> lr <- -2*(model_aft_exp$loglik[2]-model_aft_wei$loglik[2])
> pval <- pchisq(lr, df=1, lower.tail=FALSE)
> pval
[1] 0.383884
```

Appendix.

(Q4)
```
brain <- read.csv("data/brain.csv")
library(survival)
```

```
Surv(brain$weeks, brain$event)

# Get Kaplan-Meier
fit_treat <- survfit(Surv(weeks,event)~treat,
                     data = brain)

library(survminer)
ggsurvplot(
  fit_treat,
  data = brain,
  # Use NULL if survival object is provided directly
  # title = "Kaplan-Meier Survival Curve",
  xlab = "Time (weeks)",
  ylab = "Survival Probability",
 # risk.table = TRUE,  # Display a table with number at risk
 # pval = TRUE,        # Display p-value
 # surv.median.line = "hv",   #Add median survival line
  ggtheme = theme_minimal(),   #Adjust the plot theme if needed
  legend.title = ''
)+
  geom_line(data = brain_dat,
  aes(x=time,y = survivalP,color=treat),linetype='dashed')+

  ggsurvplot(
    fit_treat,
    data = brain,
    xlab = "Time (weeks)",
    ylab = "Survival Probability",
    # risk.table = TRUE,
    # pval = TRUE,
    # surv.median.line = "hv",
    ggtheme = theme_minimal(),
    legend.title = ''
  ) +
  geom_line(data = brain_dat,
  aes(x = time, y = survivalP,
  color = treat), linetype = 'dashed')


fit_pois_Q4 <- glm(event ~ treat, offset=log(weeks),
                   family=poisson(link=log), data=brain)

coef_table_Q4 <- summary(fit_pois_Q4) %>% coef() %>% data.frame()

lambda0 <- exp(coef_table_Q4$Estimate[1]) ## lambda0 = exp(intercept)
```

```
beta <- coef_table_Q4$Estimate[2] ## beta: coef in poission dist

median_ts <- c(t0,t1,km_t0,km_t1)


brain_dat <- data.frame(time = seq(0,200,0.01),
                        lambda0 = lambda0,
                        beta = beta
                       ) %>%
  mutate(S0 = exp(-(time*lambda0)),
         S1 = exp(-(time*lambda0*exp(beta)))) %>%
  tidyr::pivot_longer(cols = c("S0","S1"),
                      values_to = 'survivalP',
                      names_to = "treat")

km_curve <- ggsurvplot(
  fit_treat,
  data = brain,
  xlab = "Time (weeks)",
  ylab = "Survival Probability",
  ggtheme = theme_bw(),
  surv.median.line = "hv",
  surv.median.time = TRUE,
  legend.title = ''
)

plot1 <-  km_curve$plot +
  scale_color_manual(values = c("#DD5F20",
  "skyblue","#DD5F20", "skyblue"))+
  scale_linetype_manual(values = c( "solid",
  "solid","dashed","dashed"))+
  geom_line(data = brain_dat,
  aes(x = time, y = survivalP, color = treat),
  linetype='dashed',linewidth=0.9)


ggsave(plot1,file='km_exp_surv_curves.png',width = 7.11,height = 6.83,
       dpi=300)

## median survival time exponential model
S0 <- function(t,s){exp(-(t*lambda0))-s}
S1 <- function(t,s){exp(-(t*lambda0*exp(beta)))-s}
t0 <- uniroot(S0, s=0.5, interval=c(0,200))$root %>% round(1)
t1 <- uniroot(S1, s=0.5, interval=c(0,200))$root %>% round(1)
```

```
## median survival time KM curve
km_median_dat <- summary(fit_treat)$table %>% data.frame()
km_t0 <- km_median_dat$median[1]
km_t1 <- km_median_dat$median[2]

plot2 <-  plot1 +
  geom_segment(x = 0, y = 0.5, xend = t1, yend = 0.5,
               linetype = "dashed", color = "black") +
  geom_segment(x = t0, y = 0, xend = t0, yend = 0.5,
  linetype = "dashed", color = "#00BB00") +
  geom_segment(x = t1, y = 0, xend = t1, yend = 0.5,
  linetype = "dashed", color = "#00BB00") +
  scale_x_continuous(breaks = c(0, median_ts, 50, 100, 150, 200),
  labels = c("0", paste0(median_ts), "50", "100", "150", "200")) +
  theme(
    axis.text.x = element_text(size = 5, angle = 45, hjust = 1,
    color = c("black",  "black","#00BB00", "black",
    "#00BB00", "black", "black", "black", "black")),
    panel.grid = element_blank() # Remove grid lines
  )

ggsave(plot2,file='median_curves.png',width = 7.11,height = 6.83,
       dpi=300)



## nelson-Aelen cumulative hazard rate
nelson_cumhaz_dat <- data.frame(time = summary(fit_treat)$time,
                                cumhaz = summary(fit_treat)$cumhaz,
                                strata = summary(fit_treat)$strata) %>%
  mutate(lower = cumhaz - 1.96*summary(fit_treat)$std.chaz,
         upper = cumhaz + 1.96*summary(fit_treat)$std.chaz,)

library(ggplot2)

### exponential model cumulative hazard
## lambda = lambda0 exp(beta*x)
##
sum(brain$treat[brain$treat==1])/sum(brain$weeks[brain$treat==1])
pois_cumhaz <- data.frame(time = summary(fit_treat)$time) %>%
  mutate(
    pois_cumhaz1 = lambda0 * exp(beta * 1) * time,
    pois_cumhaz0 = lambda0 * exp(beta * 0) * time
  ) %>%
  tidyr::pivot_longer(cols = c("pois_cumhaz0","pois_cumhaz1"),
```

```
                        values_to = 'pois_cumhaz',
                        names_to = 'treat')

cumhaz_p <- nelson_cumhaz_dat %>%
  ggplot(aes(x = time, y = cumhaz, color = strata)) +
  geom_step()+
  geom_line(data =pois_cumhaz,
            aes(x = time,
            y =pois_cumhaz,color = treat ),linetype='dashed' )+
  scale_color_manual(values = c("#DD5F20",
  "skyblue","#DD5F20", "skyblue"))+
  theme_bw()+
  theme(legend.position = "bottom",
        legend.title = element_blank())+
  ylab("Cumulative Hazard Rate")

ggsave(cumhaz_p,file='cumhaz_p.png',width = 6.21,height = 5.9,dpi=300)
```

(Q5)
```
library(eha)
model_wei_ph <- phreg(Surv(weeks, event) ~ treat,
                      dist = "weibull",
                      data=brain)

model_wei_ph
model_wei_ph_dat <- summary(model_wei_ph) %>%
  coefficients() %>% data.frame()

lower <- exp(model_wei_ph_dat$coef -1.96*model_wei_ph_dat$se.coef.)
upper <- exp(model_wei_ph_dat$coef +1.96*model_wei_ph_dat$se.coef.)

model_wei_ph_dat$exp.coef. %>% round(2)
lower %>% round(2)
upper %>% round(2)

model_exp_ph <- phreg(Surv(weeks, event) ~ treat,
                      dist = "weibull",shape =1,
                      data=brain)
smmary(model_exp_ph)

lr <- -2*(model_exp_ph$loglik[2]-model_wei_ph$loglik[2])
pval <- pchisq(lr, df=1, lower.tail=FALSE)
pval
```

```
## or use AFT parametrization
## exponential model
model_aft_exp <- survreg(Surv(weeks, event) ~ treat,
                         dist="exponential", data=brain)
summary(model_aft_exp)
## weibull model
model_aft_wei <- survreg(Surv(weeks, event) ~ treat,
                         dist="weibull", data=brain)

summary(model_aft_exp)
-2*(model_aft_exp$loglik[2]-model_aft_wei$loglik[2])

lr <- -2*(model_aft_exp$loglik[2]-model_aft_wei$loglik[2])
pval <- pchisq(lr, df=1, lower.tail=FALSE)
pval
```