CHL 5209 Assignment 2

Belina Jang

Question 1

(a) Let the parameters in the model in Assignment 1 Q1(a) be denoted by α and β. Derive the restricted maximum likelihood estimator for α under H0 : $\beta = 0$, calculate it's value, and the value of the log-likelihood at this point. Fit the corresponding model via glm to verify the results.

The model in Assignment 1 Q1(a): $log(\lambda_Z) = \alpha + \beta \mathbf{Z}$ Under the $H_0 : \beta = 0$, the model becomes $log(\lambda) = \alpha$.

1-a)

$\log(\lambda) = \alpha$

$\ell(\alpha) = \sum_i [d_i \log(\lambda) - y_i\lambda - \log(d_i!)]$

$= \sum_i [d_i \alpha - y_i e^{\alpha} - \log(d_i!)]$

$\frac{\partial \ell}{\partial \alpha} = \sum_i [d_i - y_i e^{\alpha}]$

let it = 0

$0 = \sum_i^n [d_i - y_i e^{\hat{\alpha}}]$

$0 = \sum d_i - e^{\hat{\alpha}} \sum y_i$

$e^{\hat{\alpha}} = \frac{\sum d_i}{\sum y_i}$

taking log

$\hat{\alpha} = \log\left(\frac{\sum d_i}{\sum y_i}\right)$

log-likelihood value @ $\hat{\alpha}$ = $\ell(\hat{\alpha}) = \sum_i [d_i \hat{\alpha} - e^{\hat{\alpha}} - \log(d_i!)]$

$= \sum_i [d_i \cdot \log\left(\frac{\sum d_i}{\sum y_i}\right) - \frac{\sum d_i}{\sum y_i} - \log(d_i!)]$

```r
q1data <- data.frame(
  n = c(2430, 2387), #N0, N1
  d = c(45, 22), #D0, D1
  z = c(0, 1), #Z0, Z1
  y = c(3391.8, 3352.4) #Y0, Y1
)
```

```r
# calculate alpha_hat under H0: beta = 0

alpha_hat <- log(sum(q1data$d)/sum(q1data$y))
alpha_hat
```

```
[1] -4.611746
```

$\therefore \hat{\alpha} = -4.611746$.

```r
# value of the log-likelihood at alpha_hat

llh <- 0
# for loop for each row
for (i in 1:nrow(q1data)) {
  di <- q1data$d[i]
  yi <- q1data$y[i]
  zi <- q1data$z[i]
  xi <- q1data$x[i]

  lambda <- exp(alpha_hat)
  llh <- llh + di * (log(lambda) + log(yi)) - yi * lambda - log(factorial(di))
}

llh
```

```
[1] -9.188345
```

$\therefore$ The value of the log-likelihood function at the maximum likelihood point (restricted): $l(\hat{\alpha}) = -9.188345$.

```r
# verify with glm
model <- glm(d ~ 1, offset = log(y), family = poisson(link = "log"), data = q1data)
summary(model)
```

```
Call:
glm(formula = d ~ 1, family = poisson(link = "log"), data = q1data,
    offset = log(y))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.6117     0.1222  -37.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 7.792  on 1  degrees of freedom
Residual deviance: 7.792  on 1  degrees of freedom
AIC: 20.377

Number of Fisher Scoring iterations: 4
```

```
logLik(model)
```

```
'log Lik.' -9.188345 (df=1)
```

∴ The glm outputs the same values calculated above.

(b) Calculate the likelihood ratio test statistic for the null hypothesis H0 : $\beta = 0$ and a p-value for the null hypothesis, and interpret the result.

```
# model with z: log(lambda_{Z})=alpha+beta*Z
model2 <- glm(d ~ z, offset = log(y), family = poisson(link = "log"), data = q1data)
summary(model2)
```

```
Call:
glm(formula = d ~ z, family = poisson(link = "log"), data = q1data,
    offset = log(y))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)  -4.3225     0.1491 -28.996  < 2e-16 ***
z            -0.7039     0.2601  -2.706  0.00681 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  7.7920e+00  on 1  degrees of freedom
Residual deviance: -4.4409e-15  on 0  degrees of freedom
AIC: 14.585

Number of Fisher Scoring iterations: 3
```

```r
D <- 2 * (logLik(model2) - logLik(model))
p_value <- pchisq(D, df = 1, lower.tail = FALSE)

cat("LRT Statistic (D):", D, "\n")
```

```
LRT Statistic (D): 7.791952
```

```r
cat("p-value:", p_value, "\n")
```

```
p-value: 0.005247946
```

```r
# note: manual calculation for log-likelihood value at the MLE
alpha_hat <- -4.3225
beta_hat <- -0.7039

llh <- 0
# for loop for each row
for (i in 1:nrow(q1data)) {
  di <- q1data$d[i]
  yi <- q1data$y[i]
  zi <- q1data$z[i]
  xi <- q1data$x[i]

  lambda_z <- exp(alpha_hat + beta_hat * zi)
  llh <- llh + di * (log(lambda_z) + log(yi)) - yi * lambda_z - log(factorial(di))
}

cat("Unrestricted: log-likelihood at MLE is", llh, "\n")
```

```
Unrestricted: log-likelihood at MLE is -5.292369
```

$\therefore$ We reject the null hypothesis that $\beta = 0$ (p-value$<$0.05). This means that including variable Z (which is the same as not restricting beta to 0) is significant in predicting the event (death) counts.

(c) In Assignment 1 Q1(d) we calculated standard errors for the unrestricted MLEs by inverting the observed information matrix. It turns out it is easier to derive these by proceeding as if the event counts D0 and D1 were independent Poisson distributed random variables, and using the Delta method variance approximation $V[g(\hat{\theta})] \approx (g'(\theta))^2 V[\hat{\theta}]$. Verify this.

1-c) $V[g(\hat{\theta})] \approx (g'(\theta))^2 V[\hat{\theta}]$

We know $\hat{\alpha} = \log\left(\frac{D_0}{Y_0}\right)$, $\hat{\beta} = \log\left(\frac{D_1 Y_0}{D_0 Y_1}\right)$

Since D_0 & D_1 follows a poisson distribution

$\text{Var}(D_0) = \text{mean}(D_0) = M_0$ and $\text{var}(D_1) = \text{mean}(D_1) = M_1$

<because $\frac{d}{dD_0}\left[\log\left(\frac{D_0}{Y_0}\right)\right] = \frac{d}{dD_0}\left[\log D_0 - \log Y_0\right] = \frac{1}{D_0} - 0 = \frac{1}{D_0}$ >

$V[\hat{\alpha}] = V\left[\log\left(\frac{\hat{D_0}}{Y_0}\right)\right] \approx \left(\frac{1}{D_0}\right)^2 \cdot V[\hat{D_0}]$ using delta method.

$= \frac{1}{D_0^2} \times M_0 = \frac{1}{D_0^2}(\lambda_0 Y_0) = \frac{1}{D_0^2}(e^{\alpha} Y_0) = \frac{1}{45^2}\left(e^{-4.3225} \cdot (3391.8)\right)$

≈ 0.14907

Let $g(D_0, D_1) = \log\left(\frac{D_1 Y_0}{D_0 Y_1}\right)$

$V[\hat{\beta}] = V\left[\log\left(\frac{D_1 Y_0}{D_0 Y_1}\right)\right] \approx \begin{bmatrix} \frac{\partial g}{\partial D_0} & \frac{\partial g}{\partial D_1} \end{bmatrix} \begin{bmatrix} \text{Var}(D_0) & 0 \\ 0 & \text{var}(D_1) \end{bmatrix} \begin{bmatrix} \frac{\partial g}{\partial D_0} \\ \frac{\partial g}{\partial D_1} \end{bmatrix}$ ∈

$\frac{\partial g}{\partial D_0} = \frac{\partial}{\partial D_0} \log \frac{D_1 Y_0}{D_0 Y_1} = \frac{D_0 Y_1}{D_1 Y_0} \times \frac{-D_1 Y_0}{Y_1 D_0^2} = -\frac{1}{D_0}$

$\frac{\partial g}{\partial D_1} = \frac{\partial}{\partial D_1} \log \frac{D_1 Y_0}{D_0 Y_1} = \frac{D_0 Y_1}{D_1 Y_0} \times \frac{Y_0}{D_0 Y_1} = \frac{1}{D_1}$

$\ni \begin{bmatrix} -\frac{1}{D_0} & \frac{1}{D_1} \end{bmatrix} \begin{bmatrix} \text{Var}(D_0) & 0 \\ 0 & \text{var}(D_1) \end{bmatrix} \begin{bmatrix} -\frac{1}{D_0} \\ \frac{1}{D_1} \end{bmatrix}$

$= \left(-\frac{1}{D_0}\right)^2 \text{var}\, D_0 + \left(\frac{1}{D_1}\right)^2 \text{var}\, D_1$

$= \frac{\text{Var}\, D_0}{D_0^2} + \frac{\text{var}\, D_1}{D_1^2} = \frac{M_0}{D_0^2} + \frac{M_1}{D_1^2}$

$= \frac{\lambda_0 Y_0}{D_0^2} + \frac{\lambda_1 Y_1}{D_1^2} = \frac{e^{\alpha} Y_0}{D_0^2} + \frac{e^{\alpha+\beta} Y_1}{D_1^2} \approx 0.26014$

Question 2. Assume the a linear model $log(\tilde{T}i) = \alpha + \gamma xi + \sigma\epsilon$ for the log event time so that $S_0(t)$ is the survival function of the random variable $e^{\alpha+\sigma\epsilon}$.

(a) Show that above model implies the property $S_i(te^{\gamma xi}) = S_0(t)$ for survival functions, where S_i is the survival function of individual with covariate value x_i.

2-a)

$\log \tilde{T}_i = \alpha + \gamma x_i + \sigma\varepsilon$

exponentiating both sides

$\tilde{T}_i = e^{\alpha+\gamma x_i+\sigma\varepsilon}$ then $\tilde{T}_0 = e^{\alpha+\sigma\varepsilon}$

substituting $e^{\alpha+\sigma\varepsilon} = \tilde{T}_0$ back to $\tilde{T}_i$

$\tilde{T}_i = e^{\alpha+\sigma\varepsilon} \cdot e^{\gamma x_i} = \tilde{T}_0 \cdot e^{\gamma x_i}$

$S_i(t) = P(\tilde{T}_i > t) = P(\tilde{T}_0 \cdot e^{\gamma x_i} > t) = P(\tilde{T}_0 > te^{-\gamma x_i})$

$S_i(te^{\gamma x_i}) = P(\tilde{T}_0 > te^{\gamma x_i} \cdot e^{-\gamma x_i}) = P(\tilde{T}_0 > t)$

we know $S_0(t) = P(\tilde{T}_0 > t) \Rightarrow S_0(t) = P(\tilde{T}_0 > t) = S_i(te^{\gamma x_i})$ ✱

$\Rightarrow S_0(t) = S_i(te^{\gamma x_i})$ ■

(b) Show that above model implies the property $\lambda_i(t) = e^{-\gamma x_i}\lambda_0(te^{-\gamma x_i})$ for hazard functions.

2-b)

from ✱ $S_0(t) = S_i(te^{\gamma x_i})$, we also get $S_0(te^{-\gamma x_i}) = S_i(te^{-\gamma x_i} \cdot e^{\gamma x_i}) = S_i(t)$

we know $\lambda_i(t) = \frac{f_i(t)}{S_i(t)} = -\frac{d}{dt}[S_i(t)] \times \frac{1}{S_i(t)}$

$= -\frac{d}{dt}[S_0(te^{-\gamma x_i})] \times \frac{1}{S_0(te^{-\gamma x_i})}$

$= -[-f_0(te^{-\gamma x_i}) \cdot e^{-\gamma x_i}] \times \frac{1}{S_0(te^{-\gamma x_i})} = \frac{f_0(te^{-\gamma x_i})}{S_0(te^{-\gamma x_i})} \cdot e^{-\gamma x_i}$ ($= \lambda_0(te^{-\gamma x_i})$)

$= e^{-\gamma x_i} \cdot \lambda_0(te^{-\gamma x_i})$ ■

Question 3.

(a)

3. (a) The Kaplan-Meier estimator is given by

$$\hat{S}_{\mathrm{KM}}(t) = \prod_{j:t_{(j)} \le t} \left(1 - \frac{d_j}{n_j}\right),$$

where $t_{(j)}$ are the ordered event times in the sample, d_j is the number of events at time t_j, and n_j the size of the riskset at time t_j. Show that in the absence of censoring and ties, we have that $\hat{S}_{\mathrm{KM}}(t) = 1 - \hat{F}(t)$, where $\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{t_i \le t\}}$ is the empirical cumulative distribution function (ECDF).

3-a) $t_{(j)}$: ordered event times in the sample

d_j : # of events @ time t_j

n_j : size of the risk set at time t_j

Show: $\hat{S}_{KM}(t) = \prod_{j:t_{(j)} \le t} \left(1 - \frac{d_j}{n_j}\right) = 1 - \frac{1}{n}\sum 1_{\{t_i \le t\}}$

in the absense of censoring and ties, every individual had events and each event time t_j have one d_j. $\Rightarrow$ all $d_j = 1$

then $n_j = n - j + 1$ since it's ordered.

Substituting $d_j = 1$ & $n_j = n - j + 1$, we get $\hat{S}_{KM}(t) = \prod_{j:t_{(j)} \le t} \left(1 - \frac{1}{n-j+1}\right)$

$= \prod \left(\frac{n-j}{n-j+1}\right) = \frac{n-1}{n} \cdot \frac{n-2}{n-1} \cdot \frac{n-3}{n-2} \cdots \frac{n-j}{n-j+1} = \frac{n-j}{n}$

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

where j is the last j: $t_{(j)} \le t$

Since j is same as the # of $t_{(j)}$ that are $\le t$ which is the same as # of $t_{(i)} \le t$ $(= \sum 1_{\{t_i \le t\}})$

Therefore, $\hat{S}_{KM}(t) = 1 - \hat{F}(t)$ in the absense of censoring & ties. ∎

(b)

(b) The K-M estimator has pointwise variance formula

$$\hat{V}(t) = \hat{S}_{\mathrm{KM}}(t)^2 \sum_{j:t_{(j)} \le t} \frac{d_j}{n_j(n_j - d_j)}$$

known as the Greenwood formula. Show that in the absence of censoring and ties, we have that

$$\hat{V}(t) = \frac{\hat{S}_{\mathrm{KM}}(t)[1 - \hat{S}_{\mathrm{KM}}(t)]}{n} = \frac{\hat{F}(t)[1 - \hat{F}(t)]}{n}.$$

(Hint: show that $\hat{V}(t)/\hat{S}_{\mathrm{KM}}(t)^2$ and $(1-\hat{S}_{\mathrm{KM}}(t))/(n\hat{S}_{\mathrm{KM}}(t))$ are equal by observing that they are constant between the observed survival times and by showing that their increments at the observed survival times are the same.)

from 3a) we know $d_j = 1$ for all j and $n_j = n - j + 1$ in the absence of censoring and ties.

$$\hat{V}(t) = \hat{S}_{KM}(t)^2 \sum_{j:t_{(j)} \le t} \frac{d_j}{n_j(n_j - d_j)} = \hat{S}_{KM}(t)^2 \sum_{j:t_{(j)} \le t} \frac{1}{n_j(n_j - 1)}$$

① $$\frac{\hat{V}(t)}{\hat{S}_{KM}(t)^2} = \sum_{j:t_{(j)} \le t} \frac{1}{n_j(n_j-1)} = \sum_{j} \frac{1}{(n-j+1)(n-j)}$$

from part a) we know $\hat{S}_{KM}(t) = \frac{n - j_t}{n} = 1 - \frac{j_t}{n} \Rightarrow 1 - \hat{S}_{KM}(t) = \frac{j_t}{n}$

so ② $$\frac{1 - \hat{S}_{KM}(t)}{n\hat{S}_{KM}(t)} = \frac{j}{n(n-j)}$$

for ① it's easy to see the increment b/w j and $j-1$ will be $\frac{1}{(n-j+1)(n-j)}$

for ②

Increment $= \frac{j}{n(n-j)}$ (@ j) $- \frac{j-1}{n(n-(j-1))}$ (@ $j-1$) $= \frac{j}{n(n-j)} - \frac{j-1}{n(n-j+1)}$

$$= \frac{j(n-j+1) - (j-1)(n-j)}{n(n-j)(n-j+1)} = \frac{jn - j^2 + j - (jn - j^2 - n + j)}{n(n-j)(n-j+1)}$$

$$= \frac{n}{n(n-j)(n-j+1)} = \frac{1}{(n-j)(n-j+1)}$$ which is the same as increment for ①

and due to the absence of censoring & ties and every individuals had events at $t_{(j)}$ which also means there's no event between time t_j and t_{j+1} so ① & ② are constant between this period (obs' survival times)

Therefore ① = ② : $\frac{\hat{V}(t)}{\hat{S}_{KM}(t)^2} = \frac{1-\hat{S}_{KM}(t)}{n\,\hat{S}_{KM}(t)}$

$$\Rightarrow \quad \hat{V}(t) = \frac{1-\hat{S}_{KM}(t)}{n\,\hat{S}_{KM}(t)} \cdot \hat{S}_{KM}(t)^2$$

$$= \frac{(1-\hat{S}_{KM}(t))\,\hat{S}_{KM}(t)}{n} \quad ③$$

Using $\hat{S}_{KM}(t) = 1-\hat{F}(t)$ from part a,

$$③ \quad \frac{\hat{F}(t)\,(1-\hat{F}(t))}{n}$$

■

(c)

(c) The previous Greenwood formula gives the approximation

$$\hat{V}(t) \approx V\left(\hat{S}_{\mathrm{KM}}(t)\right).$$

This could be used to obtain 95% pointwise confidence bands for the survival function through $\hat{S}_{\mathrm{KM}}(t) \pm 1.96\sqrt{\hat{V}(t)}$. The problem with these confidence bands is that they are not bounded by $[0,1]$. Use the Delta method to obtain approximation for $V\left(\log(-\log \hat{S}_{\mathrm{KM}}(t))\right)$, and explain how this result can be used to obtain confidence bands that are bounded by $[0,1]$.

$$V[g(x)] \approx [g'(x)]^2 V[x]$$

let $g(x) = \log(-\log(x))$

$$g'(x) = \frac{1}{-\log x} \cdot -\frac{1}{x} = \frac{1}{x \log x}$$

$$V[g(\hat{S}_{KM}(t))] \approx [g'(S_{KM}(t))]^2 V[\hat{S}_{KM}(t)]$$

$$V[\log(-\log \hat{S}_{KM}(t))] \approx \left[\frac{1}{\hat{S}_{KM}(t) \cdot \log \hat{S}_{KM}(t)}\right]^2 V[\hat{S}_{KM}(t)]$$

< using $\hat{V}(t) \approx V(\hat{S}_{KM}(t))$ >

$$\approx \left[\frac{1}{\hat{S}_{KM}(t) \cdot \log \hat{S}_{KM}(t)}\right]^2 \hat{V}(t) = \frac{\hat{V}(t)}{(\hat{S}_{KM}(t) \cdot \log \hat{S}_{KM}(t))^2}$$

$\Rightarrow$ 95% Confidence band for $\log(-\log \hat{S}_{KM}(t))$:

$$\log(-\log \hat{S}_{KM}(t)) \pm 1.96 \sqrt{\frac{\hat{V}(t)}{(\hat{S}_{KM}(t) \cdot \log \hat{S}_{KM}(t))^2}}$$

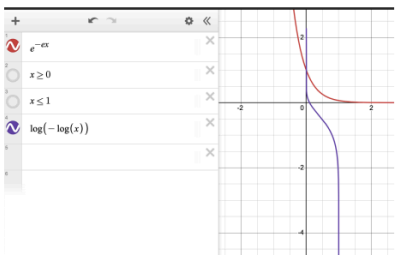
Say this equals to (lower_bd, upper_bd)

and for $\log(-\log \hat{S}_{KM}(t))$, $\hat{S}_{KM}(t)$ is bounded by $(0,1)$. (otherwise undefined)

$\Rightarrow$ 95% Confidence band for $\hat{S}_{KM}(t)$ becomes

$$\left(e^{-e^{\text{lower\_bd}}}, e^{-e^{\text{upper\_bd}}}\right)$$

$\therefore$ the transformation $e^{-e^{*}}$ ensures outcome prob are always bounded by $[0,1]$.



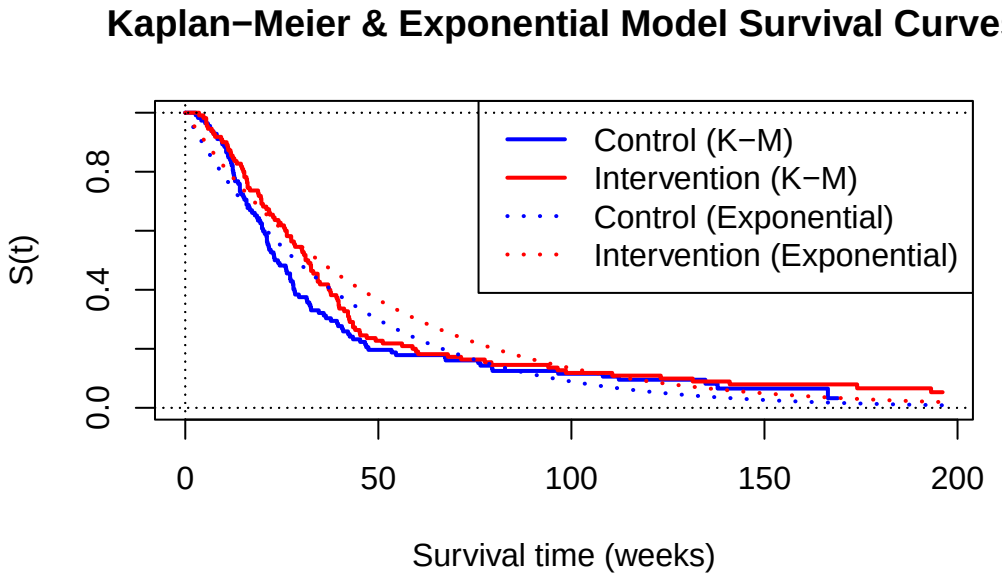
Question 4. The dataset brain.csv are from a randomized trial of a new treatment (chemotherapeutic implant placed at surgery) vs standard care (placebo implant placed at surgery) for patients with recurrent malignant brain tumours scheduled for resection who had previously underwent radiation therapy. The primary outcome is survival time from the date of randomization.

```
library(survival)
library(ggplot2)

# load the data
brain <- read.csv("brain.csv")
```

(a) Present the Kaplan-Meier survival curves in the two treatment arms, and compare these to the survival curves given by the exponential model. Based on the visual, what can you say about the fit of the exponential model?

```
# Fit Kaplan-Meier survival curves by treatment group
km_fit <- survfit(Surv(weeks, event) ~ treat, data=brain, conf.type = "plain")

plot(km_fit, col = c("blue", "red"), lwd = 2,
     xlab = "Survival time (weeks)", ylab = "S(t)",
     main = "Kaplan-Meier & Exponential Model Survival Curves")
abline(h = c(0, 1), lty = "dotted")
abline(v = c(0), lty = "dotted")

# exponential model survival curves

S0 <- function(t,r0, s = 0) {exp(-r0*t)-s}
S1 <- function(t,r1, s = 0) {exp(-r1*t)-s}

r1 <- sum(brain$event[brain$treat==1])/sum(brain$weeks[brain$treat==1])
r0 <- sum(brain$event[brain$treat==0])/sum(brain$weeks[brain$treat==0])

curve(S0(x, r0), from = 0, to = max(brain$weeks), col = "blue", add = TRUE,
      lwd = 2,lty=3, xlab = "Survival time (weeks)", ylab = "S(t)",
      main = "Exponential Model Survival Curves")
curve(S1(x, r1), from = 0, to = max(brain$weeks), col = "red", add = TRUE,
```

```
      lwd = 2,lty=3)

abline(h=c(0,1), lty='dotted')
abline(v=c(0), lty='dotted')
legend("topright",legend = c("Control (K-M)", "Intervention (K-M)",
                             "Control (Exponential)", "Intervention (Exponential)"),
       col = c("blue", "red","blue", "red"),lty = c(1,1,3,3) ,lwd = 2)
```



The fit of the exponential model is very similar to the Kaplan-Meier model with a smoother curve because of its constant hazard rate assumption.

(b) Find the median survival times (in weeks) in the two arms based on the Kaplan-Meier curves, as well as the exponential model.

```
# Median survival times from Kaplan-Meier curves
km_median <- summary(km_fit)$table[, "median"]
cat("Median survival time for S0: control (Kaplan-Meier):", km_median[1], "\n")
```

```
Median survival time for S0: control (Kaplan-Meier): 23.57
```

```r
cat("Median survival time for S1: treatment (Kaplan-Meier):", km_median[2], "\n")
```

```
Median survival time for S1: treatment (Kaplan-Meier): 31.5
```

```r
# Find median survivals for exponential model:
# log(2) / exp(exp_fit_control$coef[1])
t0 <- uniroot(function(t) S0(t,r0, s = 0.5), interval = c(0,max(brain$weeks)))$root

# log(2) / exp(exp_fit_treatment$coef[1])
t1 <- uniroot(function(t) S1(t,r1, s = 0.5), interval = c(0,max(brain$weeks)))$root

# Find 25% survivals:
t0_25 <- uniroot(S0,r0,s=0.25, interval = c(0,max(brain$weeks)))$root

t1_25 <- uniroot(S1,r1,s=0.25, interval = c(0,max(brain$weeks)))$root

cat("Median survival time for S0: control (Exponential):", t0, "\n")
```

```
Median survival time for S0: control (Exponential): 28.65913
```

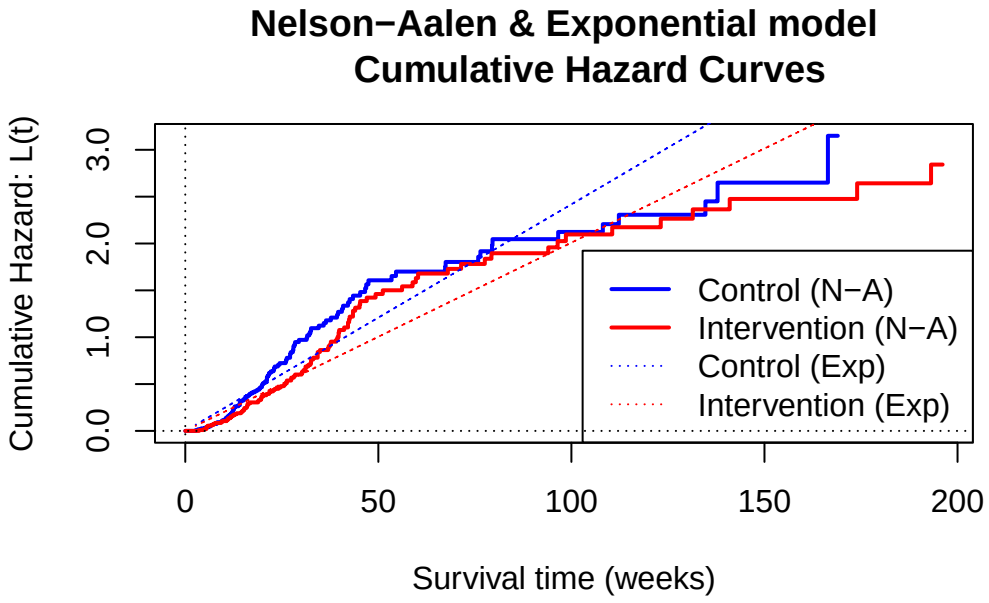
```r
cat("Median survival time for S1: treatment (Exponential):", t1, "\n")
```

```
Median survival time for S1: treatment (Exponential): 34.49379
```

(c) Present the Nelson-Aalen cumulative hazard curves in the two treatment arms, and compare these to the cumulative hazard curves given by the exponential model. Based on the visual, what can you say about the fit of the exponential model?

```r
# Nelson-Aalen curves:
plot(km_fit, col = c("blue", "red"), lwd = 2,
     xlab = "Survival time (weeks)", ylab = "Cumulative Hazard: L(t)",
     fun = "cumhaz", main = "Nelson-Aalen & Exponential model
     Cumulative Hazard Curves")
abline(h=c(0), lty='dotted')
abline(v=c(0), lty='dotted')

# exponential model curve:
```

```
lines(km_fit$time,c(r0)*km_fit$time, col = "blue", lwd = 1, lty = 3,
      xlab = "Survival time (weeks)", ylab = "Cumulative Hazard: L(t)")
lines(km_fit$time,c(r1)*km_fit$time, col = "red", lwd = 1, lty = 3)

legend("bottomright",
       legend = c("Control (N-A)", "Intervention (N-A)", "Control (Exp)",
                  "Intervention (Exp)"),
       col = c("blue", "red","blue", "red"), lty =c(1,1,3,3), lwd = c(2,2,1,1))
```



Because of the assumption of constant hazard rate for the exponential model, the fit is not as good as the Nelson-Aalen curves which allows for change in rate over time. Due to this weakness, the exponential model likely underestimated the cummulative hazard rate in the early stages and overestimated it in the later stages.

(d) Fit an appropriate Weibull regression model to the brain dataset to estimate the treatment effect hazard ratio and it's 95% confidence interval. Interpret the result. Use the likelihood ratio test to check whether the exponential model would give an adequate fit to the data instead of the Weibull model.

```
# Weibull model (PH parametrization, package eha):
```

```r
library(eha)
model1ph <- phreg(Surv(weeks, event) ~ treat, dist = "weibull", data=brain)
summary(model1ph)
```

```
Covariate             Mean       Coef     Rel.Risk   S.E.    LR p
treat                 0.544    -0.195     0.823     0.139   0.1625

Events                    207
Total time at risk        9425.7
Max. log. likelihood      -996.16
LR test statistic         1.95
Degrees of freedom        1
Overall p-value           0.162478
```

```r
loghr <- summary(model1ph)$coefficients[1]
se <- as.numeric(summary(model1ph)$coefficients[3])

cil <- exp(loghr - 1.96 * se)
ciu <- exp(loghr + 1.96 * se)

cat("Hazard Ratio:", exp(loghr), "\n")
```

```
Hazard Ratio: 0.8229051
```

```r
cat("95% CI for HR: (",cil,",",ciu,")\n")
```

```
95% CI for HR: ( 0.6260936 , 1.081584 )
```

$\therefore$ The hazard ratio for the treatment group compared to the control group is 0.8229051 (95% CI: 0.6260936, 1.081584). This means that the treatment group has an approximately 17.70949% lower hazard rate compared to the control group. But the 95% confidence interval includes 1, which means that the treatment effect may not be statistically significant ($= 0.05$).

```r
# LRT

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

D <- 2 * (logLik(model1ph) - logLik(exp_model))
```

```r
p_value <- pchisq(D, df = 1, lower.tail = FALSE)

cat("LRT Statistic (D):", D, "\n")
```

```
LRT Statistic (D): 0.7582267
```

```r
cat("p-value:", p_value, "\n")
```

```
p-value: 0.383884
```

∴ We fail to reject the null hypothesis that the exponential model provides an adequate fit to the data. (p-value>0.05). This means that the weibull model does not provide a significantly better fit to the data compared to the exponential model. So, the exponential model would give an adequate fit to the data. Therefore, choosing the simpler model (exponential model) would be better.