Hw2-Surv

Johanna Copeland

3/7/2021

Problem 1: In this problem, we step through some analyses of the trial. We are interested in examining the effect of treatment on time to death

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(eventtimedata)
library(survival)
```

```
## Warning: package 'survival' was built under R version 4.0.4
```

a) Recode the treatment variables to avoid any confusion. Either create three binary variables (one for each treatment arm), or create a single factor variable with three levels. Produce a table to confirm that the new coding is correct. Load data:

```
data(mac)
```

recode treatment

```
treatment <- rep(NA, length(mac$rif))
for (i in 1:length(treatment)){
  if (mac$rif[i] == 1){
    treatment[i] = 1
}
  if (mac$clari[i] == 1){
    treatment[i] = 2
}
  if (mac$clari[i] == 0 & mac$rif[i] == 0){
```

```
    treatment[i] = 3
}
}
mac$treatment <- treatment
table <- as.data.frame(cbind("rif" = mac$rif, "clari" = mac$clari, "treatment" = mac$treatment))

head(table, 10)
```

```
##    rif clari treatment
## 1    1     0         1
## 2    1     0         1
## 3    1     0         1
## 4    0     1         2
## 5    0     1         2
## 6    0     1         2
## 7    0     0         3
## 8    1     0         1
## 9    0     0         3
## 10   1     0         1
```

ANS: We can see from the table that my code created the proper output for the treatment groups based on the definitions given in the problem.

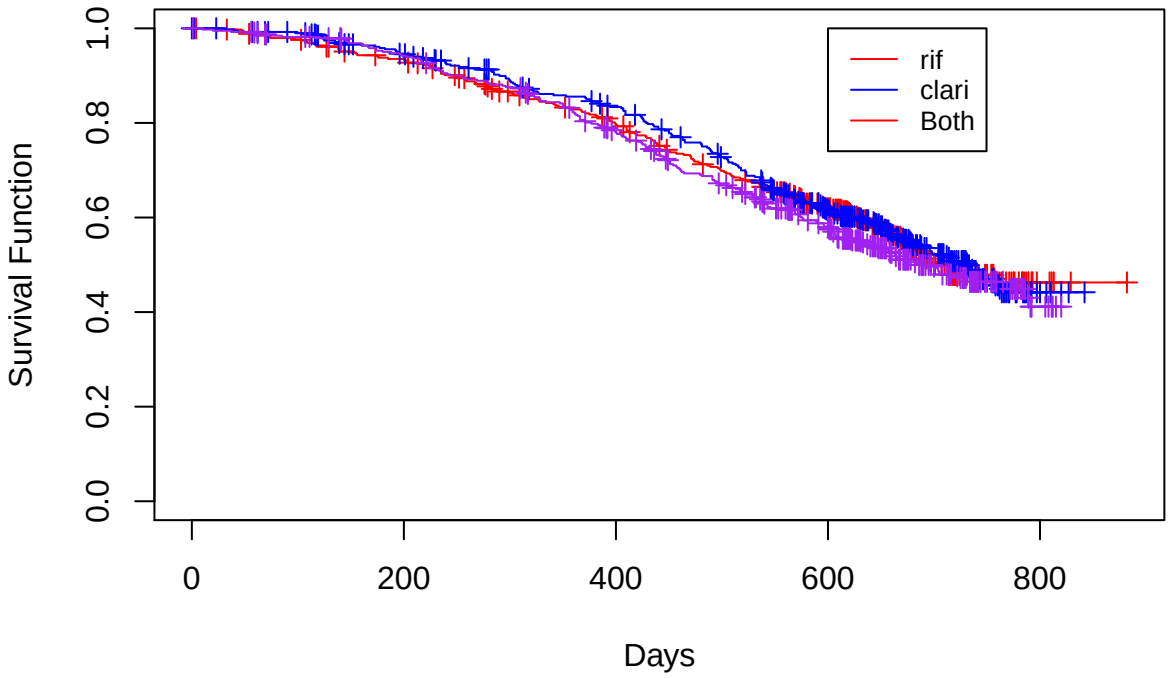
b) Explore the distribution of time to death with relevant numerical and graphical summaries by treatment. That is, estimate the median survival time and the 95% confidence interval and plot a survival curve for each treatment. Describe what you see.

```
KM.mac <- survfit(Surv(dthtime, dthstat) ~ treatment, data = mac)

KM.mac
```

```
## Call: survfit(formula = Surv(dthtime, dthstat) ~ treatment, data = mac)
##
##                n events median 0.95LCL 0.95UCL
## treatment=1 391    168    712     672      NA
## treatment=2 397    167    731     677      NA
## treatment=3 389    179    684     631      NA
```

```
plot(KM.mac, mark.time = TRUE, xlab = "Days", ylab = "Survival Function", col=c("red", "blue", "purple"
legend(x = 600, y = 1, legend=c("rif", "clari", "Both"), col=c("red", "blue"), lty=1, cex=0.8)
```



The median survival times for the rif treatment is 712 days, for clari treatment it is 731 days and both is 684 days. The confidence interval for treatment 1 is (672, NA) days, for clari it is (677, NA) days, and for both it is (631, NA) days. The reason why there is NA in the upper bounds is because it is out of scope of the study duration, meaning the event did not occur within the given timeframe. The three survival curves are plotted on the same graph and appear to be overlapping and very similar.

c) Using a proportional hazards model, calculate an overall test statistic for differences in time to death among the three treatments, without adjusting for any other covariates. Summarize your findings. Be sure to include a statement of what the null and alternative hypotheses are for the test.

```
coxph(Surv(dthtime, dthstat) ~ as.factor(treatment), data = mac)
```

```
## Call:
## coxph(formula = Surv(dthtime, dthstat) ~ as.factor(treatment),
##     data = mac)
##
##                             coef exp(coef) se(coef)      z     p
## as.factor(treatment)2 -0.02778   0.97260  0.10928 -0.254 0.799
## as.factor(treatment)3  0.08721   1.09112  0.10744  0.812 0.417
##
## Likelihood ratio test=1.25  on 2 df, p=0.5365
## n= 1177, number of events= 514
```

The overall test statistic for difference in time to death among the three treatments is 0.67 with a p-value of 0.4143. H0: beta's = 0. The survival curves are the same, meaning the treatments do not differ Ha: not

all beta's =! 0. At least one beta does not = 0 which means they are not all the same and at least one treatment differs from the baseline treatment (rif).

We see that since the pvalue is > 0.05, we fail to reject the null and can conclude that there's enough evidence for = 0, in that there is no difference in the treatment groups.

d) Repeat the analysis in part (c) using a three sample log-rank test. In this approach, what do the p-value and test statistic for differences among the three treatments correspond to in the analysis from part (c)?

```
#log rank test
KM <- survdiff(Surv(dthtime, dthstat) ~ as.factor(treatment), data = mac)
KM
```

```
## Call:
## survdiff(formula = Surv(dthtime, dthstat) ~ as.factor(treatment),
##     data = mac)
##
##                            N Observed Expected (O-E)^2/E (O-E)^2/V
## as.factor(treatment)=1 391      168      171    0.0681     0.102
## as.factor(treatment)=2 397      167      175    0.3829     0.582
## as.factor(treatment)=3 389      179      167    0.8050     1.196
##
##  Chisq= 1.3  on 2 degrees of freedom, p= 0.5
```

ANS: The log rank is referring to if the three survival curves are the same. In part c, we were looking to see if the treatments had an effect on the survival curve when comparing to the rif treatment group. From the output above, we notice that since the p-value > 0.05, we fail to reject the null to conclude that there is enough evidence to show that the 3 survival curves are the same.

e) What is the estimated survival rate at 230 days for each treatment group?

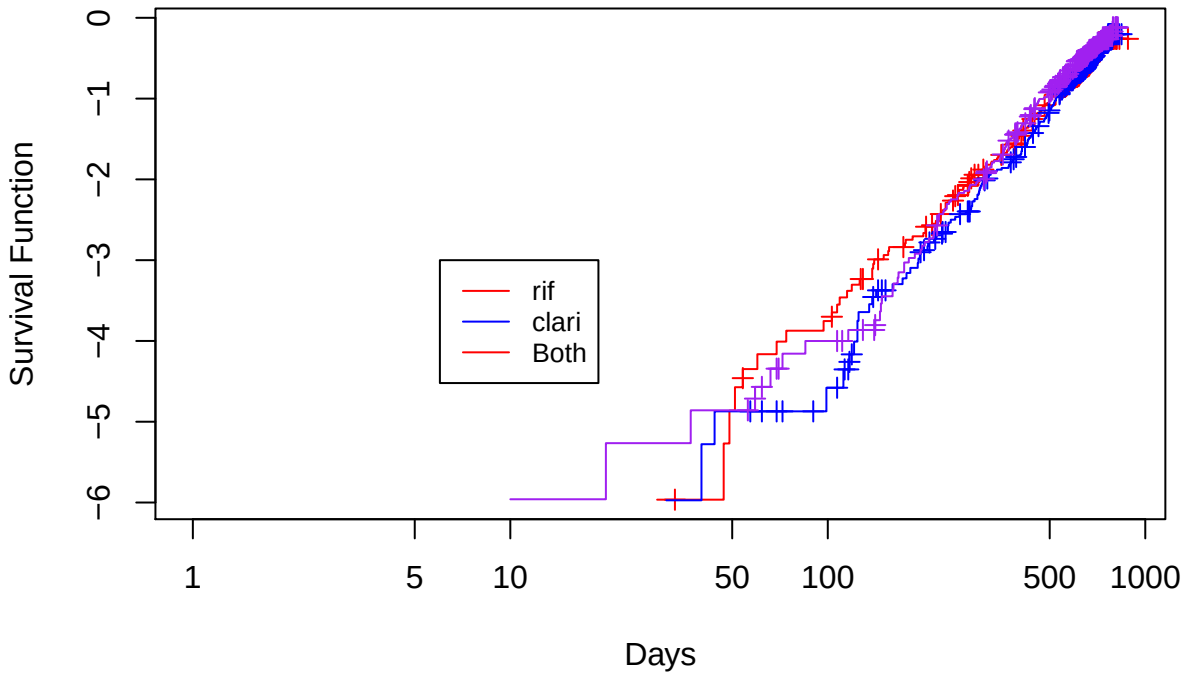
```
(est.surv = summary(KM.mac, time = 230))
```

```
## Call: survfit(formula = Surv(dthtime, dthstat) ~ treatment, data = mac)
##
##                 treatment=1
##         time       n.risk      n.event     survival      std.err lower 95% CI
##     230.0000     347.0000      34.0000       0.9117       0.0145       0.8838
## upper 95% CI
##       0.9405
##
##                 treatment=2
##         time       n.risk      n.event     survival      std.err lower 95% CI
##     230.0000     348.0000      26.0000       0.9319       0.0129       0.9069
## upper 95% CI
##       0.9575
##
##                 treatment=3
##         time       n.risk      n.event     survival      std.err lower 95% CI
##     230.0000     345.0000      32.0000       0.9157       0.0143       0.8882
## upper 95% CI
##       0.9441
```

The estimated survival rate at 230 days for rif treatment is 0.9117, For clari treatment it is 0.9319 and for both it is 0.9157.

f) Assess the assumption of proportional hazards for the three treatment groups by creating a plot of log[-log(st)] for each of the three treatments by adding fun = "cloglog" in the plot.survfit function. How should these plots look if the the proportional hazards assumption is approximately correct? Is the proportional hazards assumption valid?

```
plot(KM.mac, mark.time = TRUE, xlab = "Days", ylab = "Survival Function", col=c("red", "blue", "purple"
legend(x = 6, y = -3, legend=c("rif", "clari", "Both"), col=c("red", "blue"), lty=1, cex=0.8)
```



ANS: It is violated here since they overlap and appear to converge as the days(x-axis) increases. If the protional hazards assumption held, the survival curves would not overlap or cross over.

Problem 2: a) Fit a Cox PH model to time to death including the following variables: age, sex, karnof, antiret, cd4cat, and your treatment variable(s). Present the outputs.

```
COX2 <- coxph(formula =Surv(dthtime, dthstat) ~ age + sex + karnof + antiret + cd4cat + as.factor(treatr

COX2
```

```
## Call:
## coxph(formula = Surv(dthtime, dthstat) ~ age + sex + karnof +
##     antiret + cd4cat + as.factor(treatment), data = mac)
##
##                             coef exp(coef)  se(coef)      z        p
```

```
## age                      0.021607  1.021842  0.004994  4.327 1.51e-05
## sex                      0.330084  1.391086  0.146212  2.258   0.0240
## karnof                  -0.037386  0.963304  0.005134 -7.281 3.30e-13
## antiret                 -0.214394  0.807030  0.099682 -2.151   0.0315
## cd4cat                  -0.557419  0.572685  0.091240 -6.109 1.00e-09
## as.factor(treatment)2 -0.032853  0.967681  0.109413 -0.300   0.7640
## as.factor(treatment)3  0.022139  1.022386  0.108012  0.205   0.8376
##
## Likelihood ratio test=137.4  on 7 df, p=< 2.2e-16
## n= 1177, number of events= 514
```

```
min(mac$age)
```

```
## [1] 12
```

b) How is the "baseline" group defined in this model, in terms of the covariates? Does the baseline group correspond to any of the observations in the dataset? ANS: The baseline group in this model is 0 year old male, 0 score on karnof, no antiretroviral use, and a less than 25 Cd4 cell count. This is because it corresponds with the baseline hazard function, as in all the covariates are equal to zero. This does not correspond to any of the observations in this dataset since the minimum age in this dataset is 12.

c) Do the results from this model change the earlier conclusion about the possible differences among the three treatments? ANS: Since the p-values for the treatment groups were still > 0.05, we can say that the treatment does not have an effect on the model and our results are not different.

d) Fit the Cox PH model again after excluding your treatment variable(s). Use the likelihood ratio test statistics to examine whether treatment is significantly associated with time to death.

```
COX2.d <- coxph(formula =Surv(dthtime, dthstat) ~ age + sex + karnof + antiret + cd4cat, data = mac)

COX2.d
```

```
## Call:
## coxph(formula = Surv(dthtime, dthstat) ~ age + sex + karnof +
##     antiret + cd4cat, data = mac)
##
##              coef exp(coef)  se(coef)      z        p
## age      0.021745  1.021984  0.004986  4.362 1.29e-05
## sex      0.332205  1.394038  0.146062  2.274   0.0229
## karnof  -0.037516  0.963179  0.005128 -7.315 2.57e-13
## antiret -0.214575  0.806884  0.099521 -2.156   0.0311
## cd4cat  -0.558091  0.572300  0.091230 -6.117 9.51e-10
##
## Likelihood ratio test=137.2  on 5 df, p=< 2.2e-16
## n= 1177, number of events= 514
```

```
anova(COX2, COX2.d)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(dthtime, dthstat)
##  Model 1: ~ age + sex + karnof + antiret + cd4cat + as.factor(treatment)
##  Model 2: ~ age + sex + karnof + antiret + cd4cat
##    loglik Chisq Df P(>|Chi|)
## 1 -3324.2
## 2 -3324.3 0.262  2    0.8772
```

The p-value is greater than 0.05 so we can say that treatment is not significantly associated with death time and therefore can be dropped from the model.

e) What is the estimated hazard ratio for death associated with a higher CD4 count,adjusting for all other covariates? Give a 95% confidence interval for the hazard ratio for cd4cat, adjusting for the other covariates, and provide a verbal interpretation of the confidence interval for a non-statistician.

```
COX2
```

```
## Call:
## coxph(formula = Surv(dthtime, dthstat) ~ age + sex + karnof +
##     antiret + cd4cat + as.factor(treatment), data = mac)
##
##                             coef exp(coef)  se(coef)      z        p
## age                     0.021607  1.021842  0.004994  4.327 1.51e-05
## sex                     0.330084  1.391086  0.146212  2.258   0.0240
## karnof                 -0.037386  0.963304  0.005134 -7.281 3.30e-13
## antiret                -0.214394  0.807030  0.099682 -2.151   0.0315
## cd4cat                 -0.557419  0.572685  0.091240 -6.109 1.00e-09
## as.factor(treatment)2 -0.032853  0.967681  0.109413 -0.300   0.7640
## as.factor(treatment)3  0.022139  1.022386  0.108012  0.205   0.8376
##
## Likelihood ratio test=137.4  on 7 df, p=< 2.2e-16
## n= 1177, number of events= 514
```

```
summary(COX2)$conf.int
```

```
##                       exp(coef) exp(-coef) lower .95 upper .95
## age                   1.0218423  0.9786246 1.0118898 1.0318927
## sex                   1.3910855  0.7188631 1.0444752 1.8527190
## karnof                0.9633042  1.0380937 0.9536587 0.9730473
## antiret               0.8070300  1.2391113 0.6638051 0.9811575
## cd4cat                0.5726850  1.7461606 0.4789082 0.6848245
## as.factor(treatment)2 0.9676809  1.0333985 0.7809083 1.1991246
## as.factor(treatment)3 1.0223858  0.9781044 0.8273224 1.2634406
```

The estimated hazard ratio for cd4cat when adjusting for all other covariates is 0.5726. The 95% confidence interval is 0.4789082 to 0.6848245. Since the interval does not contain one, we can say with 95% confidence that cd4cat is not statistically significantly associated with death.

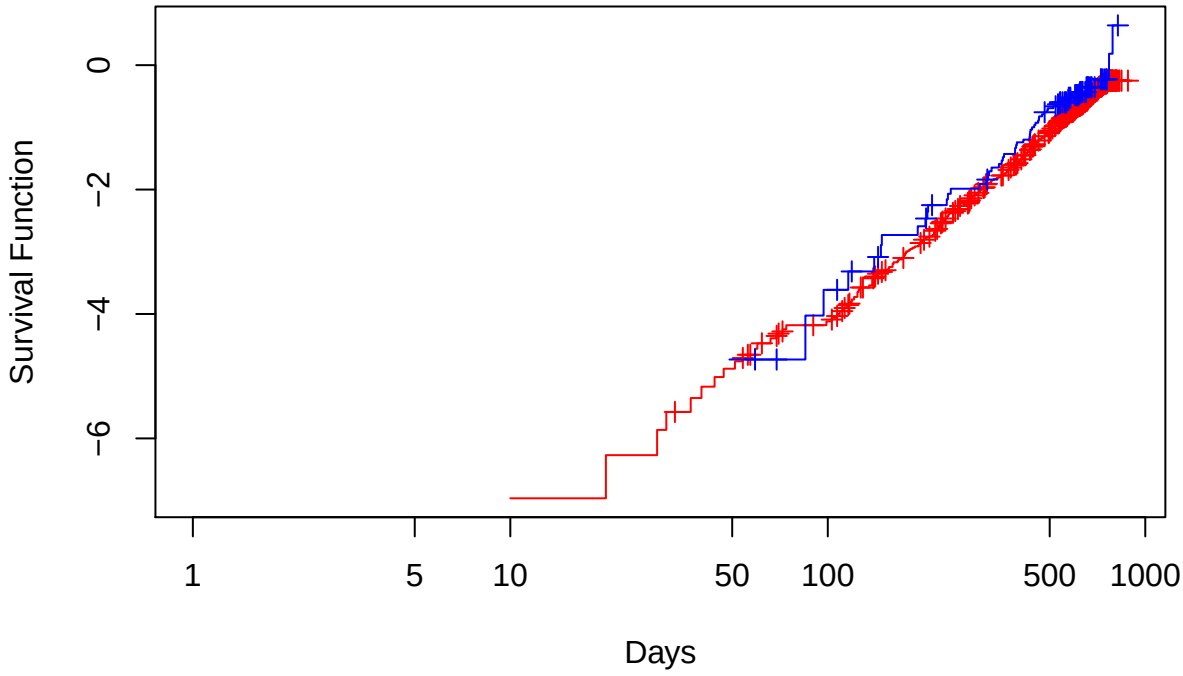
f) What is the interpretation of the estimated hazard ratio for age? What is the estimated hazard ratio for death for a subject aged 45 years versus a subject aged 30 years, holding all other covariates constant? ANS: Since the estimated hazard ratio for age is 0.9786246, that means that age is statistically associated with death when all other covariates are accounted for. Approcimately 2.2% increase per 1 year of age so for a person aged 45 years, we would say their estimated hazard ratio is .9 and 30 year old adult is about .6.

Problem 3: a) Fit a Cox PH model that includes only the variable sex. Plot log[-log(st)] to examine whether the PH assumption holds for sex.

```r
COX3 <- coxph(formula = Surv(dthtime, dthstat) ~ sex, data = mac)

sex_df <- with(mac,
               data.frame(sex = c(0, 1)))
KM3 <- survfit(Surv(dthtime, dthstat) ~ sex, data = mac)

plot(KM3, mark.time = TRUE, xlab = "Days", col = c("red", "blue"), ylab = "Survival Function", fun = "cl
```

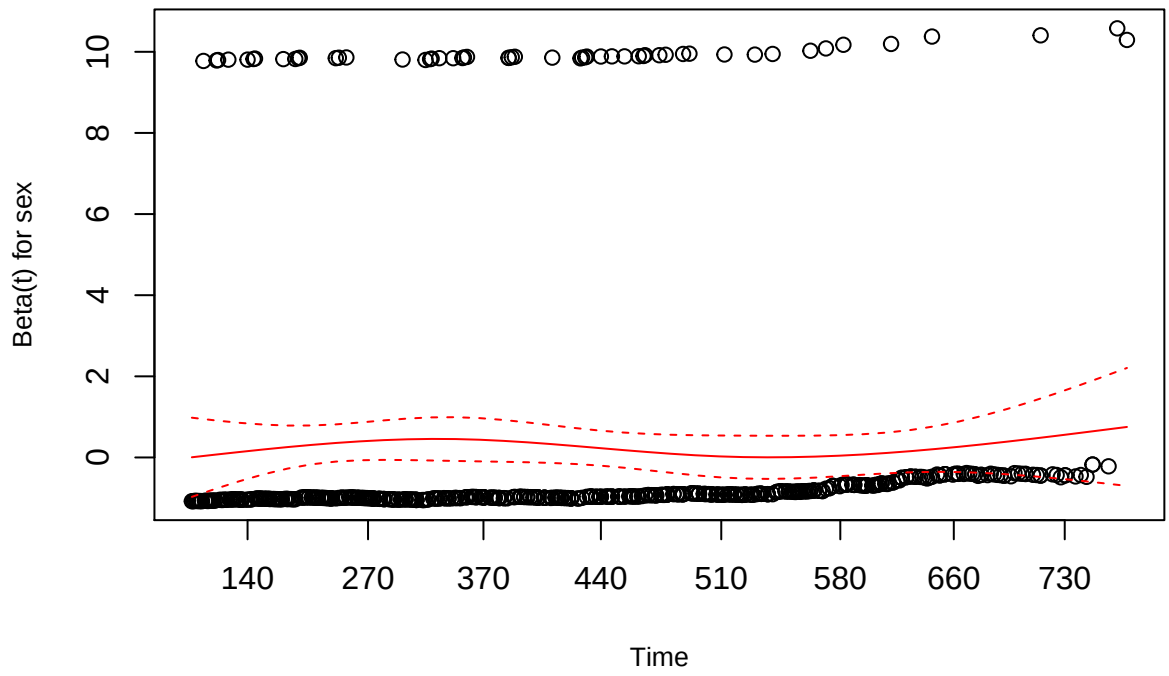


The PH assumptions do not hold since the curves cross over.

b) Use the scaled Schoenfeld residuals to examine the proportional hazards assumption for sex. Does this approach suggest the same conclusion as the one from part (a)?

```r
mac.zph = cox.zph(COX3)


plot(mac.zph, col = "red",
cex = 0.6, cex.lab = 0.8, cex.main = 0.8)
```



ANS: The proportional hazards assumption holds because the red line appears to be horizontal.

c) Use the cox.zph function to explore the PH assumption for all the variables in the full model fit from Problem 2, and describe the results.

```
macfull.zph = cox.zph(COX2)
macfull.zph
```

```
##                          chisq df       p
## age                    0.00449  1 0.94657
## sex                    0.01795  1 0.89342
## karnof                11.09532  1 0.00087
## antiret                0.68393  1 0.40824
## cd4cat                 1.52981  1 0.21614
## as.factor(treatment)   1.87925  2 0.39077
## GLOBAL                14.23916  7 0.04709
```

ANS: We see that the p-value is less than 0.05, we can say that the proportional hazard assumptions was violated. From the other p-values, it is violated for the karnof variable which has a very small p-value.