Math 473 HW 7 Spring 2023. Due Friday, March 31.

1) 2.20: The lung cancer data is the same as that described in HW5.1, but the PH model is stratified on *sex* with variables *ph.ecog*, *ph.karno*, *pat.karno* and *wt.loss*.

a) Copy and paste commands from (http://parker.ad.siu.edu/Olive/survhw.txt) for this problem into *R*. Click on the left window and hit *Enter*. Then 4 plots should appear. Include the plot in *Word*.

b) The plots are of x_j vs the martingale residuals when x_j is omitted. The loess curve should be roughly linear (or at least not taking on some simple shape such as a quadratic) if x_j is the correct functional form. If the loess curve looks like $t(x_j)$ for some simple t (eg $t(x_j) = x_j^2$), then $t(x_j)$ should be used instead of x_j. Are the loess curves in the 4 plots roughly linear?

c) Copy and paste commands from (http://parker.ad.siu.edu/Olive/survhw.txt) for this problem into *R*. Click on the left window and hit *Enter*. Then 4 plots should appear. Include the plot in *Word*. Also include the output from *cox.zph(lungfit2)* in *Word*.

d) The plots are of survival times vs scaled Schoenfeld residuals for each of the 4 variables. The loess curves should be approximately horizontal (0 slope) lines if the PH assumption is reasonable. Alternatively, the pvalue for Ho slope = 0 from *cox.zph* should be greater than 0.05 for each of the 4 variables. Is the PH assumption is reasonable? Explain briefly.

2) 2.21: a) Copy and paste the *R* commands for this problem into *R*. This problem shows how to do backward elimination for the PH model in *R* using the Leemis (1995, p. 249-250) and Lawless (1982, p. 286) lung survival data. List the AIC for the model chosen in each step. Some entries are below.

```
                                    model  AIC
perf, age, ttoent, size, type, ttype, trt  189.22  full model
perf, age, ttoent, size,       ttype, trt  187.22
.
.
.
perf,                          ttype       181.52
perf                                       183.12
```

b) Copy and paste the *R* commands for this problem into *R*. Now forward selection is done. List the AIC for the model chosen in each step. Forward selection stops at the I_{min} model. Some entries (maybe all), are listed below.

```
                                    model  AIC
perf                                       183.12
perf,                          ttype       181.52
```

3) 2.22: Copy and paste the R command

```
source("http://parker.ad.siu.edu/Olive/survpack.txt")
```

from near the top of (http://parker.ad.siu.edu/Olive/survhw.txt) into R. **(Do not give any plots for this problem.)**

For the Kaplan Meier estimator, some versions of R will give 3 curves of circles. The middle curve is the Kaplan Meier estimator while the outer two curves are the pointwise CI bands.

a) In R, type "library(survival)" if necessary. Then type "phsim(k=1)". Hit the up arrow to repeat this command several times. Repeat for "phsim(k=0.5)" and "" to make ET plots. The simulated data follows a PH Weibull regression model with $h_0(t) = kt^{k-1}$. For $k = 1$ the data follows a PH exponential regression model. Did the survival times decrease rapidly as ESP increases?

b) The function `phsim2` slices the ESP into 9 groups and computes the Kaplan Meier estimator for each group. If the PH model is reasonable and n is large enough, the 9 plots should have approximately the same shape. Type "phsim2(n=100,k=1)", then "phsim2(n=200,k=1)" and keep increasing n by 100 until the nine plots look similar (assuming survival decreases from 1 to 0, and ignoring the labels on the horizontal axis and the + signs that correspond to censored times). We will say that the plots look similar if $n = 800$. What value of n did you get?

c) The function `bphsim3` makes the slice survival plots when the single covariate is an indicator for 2 groups. The PH assumption is reasonable if the plotted circles corresponding to the Kaplan Meier estimator track the solid line corresponding to the PH estimated survival function. Type "bphsim3(n=10,k=1)" and repeat several times (use the up arrow). Do the plotted circle track the solid line fairly well?

d) The function `phsim5` is similar but the ESP takes on many values and is divided into 9 groups. Type "phsim5(n=50,k=1)", then "phsim5(n=60,k=1)" and keep increasing n by 10 until the circles track the solid lines well. We will say that the circles track the solid lines well if they are within or not very far outside the pointwise CI bands. What value of n do you get?

4) A shortcut in R for the change in PLRT test is to use the `anova` function.

```
library(survival)
source("http://parker.ad.siu.edu/Olive/survdata.txt")
alung<-as.data.frame(alung)
full <- coxph(Surv(alung[,1],alung[,2])~perf+age+ttoent+size+type+ttype+trt,
data=alung)  #full model
red <- coxph(formula = Surv(alung[, 1], alung[, 2]) ~ perf  +
    size + ttype + trt, data = alung) #reduced model
anova(full,red)
Analysis of Deviance Table
 Cox model: response is  Surv(alung[, 1], alung[, 2])
```

```
 Model 1: ~ perf + age + ttoent + size + type + ttype + trt
 Model 2: ~ perf + size + ttype + trt
   loglik  Chisq Df P(>|Chi|)
1 -87.608
2 -87.817 0.4189  3    0.9363
1-pchisq(0.42,3) #0.936
   loglik  Chisq     Df P(>|Chi|)
1
2          X^2(R|F)     pval
```

The above output shows how to use the `anova` function. To do the change in PLR test from 24) in the Exam 2 review. The value under Chisq is the test statistic $X^2(R|F)$ and the value under $P(> |Chi|)$ is the pvalue for the test. Use this information to do the four step test for whether the reduced model is good for the above output corresponding to the alung data. (The output is given, you do not need to use R).