Biostatistics 515/518, Winter 2020
HW8 (6 problems)

Use the tumor data from the clinical trial on bladder cancer discussed briefly in class when reviewing censored data.

1. Make Kaplan-Meier Curves for the placebo and treatment groups separately. Perform a logrank test and summarize the results.

2. a. Fit a Cox proportional hazards regression with "Group" as the predictor. Summarize results using language suitable for a scientific publication.
b. Compare the p-value for the regression coefficient in 2a and the pvalue from the logrank test in Q1. Are they similar? Comment.

3. Create a new variable GX in the data frame that is 1 for patients randomized to placebo and 0 for patient randomized to thiotepa. Consider fitting a simple Cox model using GX compared to the model you fit in 2 using Group. How do you think the regression parameter estimates and hazard ratios will be related? Why? Verify your predictions by fitting the Cox model using GX as the predictor.

Use the PBC dataset available at http://www.biostat.ucsf.edu/vgsm/data.html . The course text provides background on these data on pages 210-211 (DPCA Study of Primary Biliary Cirrhosis).

4. Fit a Cox PH model with albumin as the predictor. Summarize the results using language suitable for a scientific publication.

5. a. Using the fitted model, calculate the difference in log hazard for values of albumin = 2.5 g/dl, 3.5 g/dl, and 4.0 g/dl, compared to albumin=3 g/dl. Also calculate the hazard ratio of values of albumin = 2.5 g/dl, 3.5 g/dl, and 4.0 g/dl compared to albumin=3 g/dl.
b. How does the HR for 2.5 g/dl vs to 3.0 g/dl compare to the hazard ratio for 3.5 g/dl vs to 3.0 g/dl ?
c. Based on these calculations, what can you say about the estimated hazard ratio comparing 3.5 g/dl to 2.5 g/dl?

It is OK to omit confidence intervals for #5 as the purpose of the question is to help you become comfortable with Cox proportional hazards regression.

6. The Cardiovascular Health Study is a cross-sectional study of American adults aged 65 years and older. The study examined the incidence of cardiovascular disease in the elderly over an 11 year period, and a primary goal was to relate the incidence of disease to various risk factors. Within the study period, many subjects died, although many were alive at the end of the study. A few of the variables in the dataset are:

- ***age*** = Participant age at study enrollment (years)
- ***smoker*** = Indicator that the participant smokes (0 = no, 1= yes)
  ttodth = The total time (in days) that the participant was observed on study between the date of study enrollment and either death or data analysis, whichever came first.
- ***death=*** An indicator that the participant was observed to die while on study. If death=1, the number of days recorded in ttodth is the number of days between that participant's enrollment and his/her death. If death=0, the number of days recorded in ttodth is the number of days observed until the end of the study.

A question of particular interest is the association between smoking and death in this population. You decide to use Cox Proportional Hazards Regression to evaluate the association between smoking and death in this population. Notice that (ttodth, death) contain time-to-death or time-to-censoring in the standard notation.

a. Will you adjust for age in your analysis? Why or Why Not? In other words, does the crude hazard ratio for smoking or an age-adjusted hazard ratio best address the scientific question? Explain your answer.

b. Consider the output from statistical software on these data on the next page. Note the substantial difference in the estimated hazard ratio for **smoker** between the simple and the multiple regression models. Using the output provided and your scientific reasoning, explain the difference.

c. Notice that the point estimates for the hazard ratios in the software output is not the midpoint of the 95% confidence interval. Why not?

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      smoker |      4994    .1209451    .3260962          0          1
         age |      5000     72.8304    5.596418         65        100


. tabstat age, by(smoker) stat(n mean sd min q max) col(stat)

Summary for variables: age
     by categories of: smoker
smoker |   N      mean    sd    min   p25     p50    p75     max
-------+--------------------------------------------------------
     0 |4390  73.08656  5.66     65    69      72     77     100
     1 | 604  70.96689  4.68     65    67      70     73      90
-------+--------------------------------------------------------
 Total |4994   72.8302  5.60     65    68      72     76     100


. stcox smoker, robust

Cox regression -- Breslow method for ties
No. of subjects      =         4994    Number of obs   =      4994
No. of failures      =         1121
Time at risk         =     11829230
                                          Wald chi2(1)    =     22.74
Log pseudolikelihood =   -9277.4769       Prob > chi2     =    0.0000

------------------------------------------------------------------------
       |               Robust
    _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------
smoker |   1.465455   .1174366     4.77   0.000     1.252449    1.714687
------------------------------------------------------------------------

. stcox smoker age, robust
Cox regression -- Breslow method for ties

No. of subjects      =         4994         Number of obs   =      4994
No. of failures      =         1121
Time at risk         =     11829230
                                            Wald chi2(2)    =    591.94
Log pseudolikelihood =   -9033.7817         Prob > chi2     =    0.0000

------------------------------------------------------------------------
       |               Robust
    _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------
smoker |   1.924187   .1547567     8.14   0.000     1.643568    2.252719
   age |   1.114826   .0050863    23.82   0.000     1.104901     1.12484
```