BIOS 845 Homework #3

Date assigned: 04/04/2018
Due Date: 04/18/2018 by 11:59 pm (Blackboard clock time);

Instructions:

1. To receive full credit, show all work. Please make your work legible.
2. Total points for this homework are 100.
3. Do not forget to write your name on the homework.
4. Insert page numbers on all pages and also total # of pages submitted.
5. Homework can be typed or hand-written. Provide SAS code wherever necessary.
6. Use the BLACKBOARD drop box to turn in the homework (preferably as pdf) or bring it to class on 04/18/2018.

Question # 1: **36 points**

This assignment will use the data set WHAS.SAS7BDAT that can be downloaded from Blackboard. These data come from the Worcester Heart Attack Study (WHAS) and represent survival data for 481 subjects following hospital admission for acute myocardial infarction. The data set contains the following variables:

ID = Identification Code (1-481)
AGE = Age (years)
AGE_CAT = Categorized Age: 1 = [24,60), 2 = [60,69), 3 = [69, 78), 4 = 78+
SEX = Gender (1 = Male, 0 = Female)
LENFOL = Total length of follow-up (days)
FSTAT = Status as of last follow-up (1 = dead, 0 = alive)

A. Using all data from the WHAS, with length of follow-up as the survival time variable and status at last follow-up as the censoring variable, fit the proportional hazards model containing AGE, SEX and AGE*SEX interaction. Find the most parsimonious model and interpret the hazard ratio(s) from your model.

B. Using the full model from Part-A above, estimate the hazard ratio for SEX at age 50, 60, 65, 70 and 80. Also, estimate the hazard ratio for a 10-year increase in age for each gender.

C. Using the full model from Part-A above, compute and graph the estimated survival functions for 65 year old males and females and estimate the median survival times.

D. Consider now the categorized version of AGE in place of the continuous AGE and repeat the analyses done in Part-A. Then estimate the hazard ratio for SEX in each age group.

E. Suppose that you are helping a researcher with an analysis that will be written and submitted for publication to a medical journal. Would you prefer to report the results from the model in Part-A or Part-D? Defend your answer. For the model of your choice, present the results in a table suitable for publication. The table should contain point estimates along with corresponding confidence intervals.

F. Using the various methods that you learned in your class, assess the validity of the PH assumption for SEX and AGE (continuous). You may ignore the interaction term for this analysis.

G. Again, consider the model in Part-F. Using the various diagnostic checking methods that you learned in your class assess the adequacy of the model fit, check for outliers and influential observations, and assess if the function form of AGE is appropriate.

Question # 2: **12 points**

This problem will use the data set PNEUMONIA.SAS7BDAT that can be downloaded from the course website. These data consist of observations on 46 patients regarding recovery time from walking pneumonia. The purpose of the study was to determine whether there is a relationship between the initial dosage level of a drug and the time to recovery. The response variable is the number of days until recovery or censoring. Since older patients typically take longer to recover, information on age of each participant was collected as a possible covariate. The data set contains the following variables:

DRUG = Initial dosage level (50 or 100) TIME = Recovery time (days)
AGE = Age at entry (years) CURED = Status (1=Recovery, 0=Censored)

A. Treating the variable DRUG as a *continuous* covariate, fit a model which assumes proportional hazards to examine the effect of initial dosage level (drug) on recovery time, adjusting for age. Is this effect significant at the 0.05 significance level? Report a point estimate for this effect. How can this point estimate be interpreted.

B. A colleague points out that the effect of the drug may have more to due with the actual level of the drug remaining in the body at a given time than the initial dosage given. In other words, as the level of the drug in the blood system decreases, the drug's effectiveness will diminish. If it is known that the half-life of the drug is close to two days, then we can say that the actual concentration level of the drug in the patient's blood equals the initial dosage times exp(-0.35*t*), where *t* represents the time point of interest. For example, if a person is given 100mg of the drug, then at 5 days after treatment the amount remaining in his/her system is equal to

$$100 \times \exp(-0.35 \times 5) = 17.38 \text{ mg}$$

Fit a model, similar to Part-A, but model the effect of drug as the amount of drug remaining in the body at a given time rather than the initial dosage. Is the effect of drug remaining in the body significant at the 0.05 level? Report a point estimate for this effect. How can this point estimate be interpreted?

C. Note that the model that you have been asked to fit in Part-B assumes that the effect of drug remaining in the body over a given time is constant over time. However, since the amount of drug in the body decreases in a nonlinear fashion the hazard ratio comparing two individuals receiving different initial dosages may change over time. Fit a model that captures this added complexity. Then based on your model, compute the hazard ratio for comparing the hazard for recovery of a patient with an initial dosage level of 100 mg vs. a patient with an initial dosage level of 50 mg at 1 week (day 7) and 2 weeks (day 14).

Question # 3: **20 points**

This problem will use the data set LEUKEMIAB.SAS7BDAT that we have used previously during the class. Recall that this data set consists of remission survival times on 42 leukemia patients, half of whom receive a new therapy and the other half of whom get a standard therapy. The data set contains the following variables:

WEEKS = Time to Relapse (weeks) RELAPSE = Status (1=Relapse, 0=Censored)
RX = Treatment Group (1=Standard, 0=New) SEX = Gender (1=Male, 0=Female)
LOGWBC = Log White Blood Cell Count

A. In class, we determined that the RX and LOGWBC variables appeared to satisfy the proportional hazards assumption. However, we were concerned that the SEX variable might not satisfy the PH assumption. We considered several alternatives in class that allowed for the effect of SEX to not satisfy this assumption. We will now consider an alternate approach to controlling for SEX using an extended Cox model. We divide the time axis into five time intervals of four weeks duration, and define a single time dependent covariate that increases linearly with time according to the following formula: For the situation just described, write down the extended Cox model, in terms of the hazard function, which contains RX, LOGWBC, and allows for the effect of SEX to vary over time in the manner just described.

$$\text{SEX} \times \text{TIME} = \begin{cases} \text{SEX} \times 1 & \text{if } t < 4 \text{ weeks} \\ \text{SEX} \times 3 & \text{if } 4 \leq t < 8 \text{ weeks} \\ \text{SEX} \times 5 & \text{if } 8 \leq t < 12 \text{ weeks} \\ \text{SEX} \times 7 & \text{if } 12 \leq t < 16 \text{ weeks} \\ \text{SEX} \times 9 & \text{if } t \geq 16 \text{ weeks} \end{cases}$$

B. Using the model described in Part-A, express the hazard ratio for the effect of SEX adjusted for RX and LOGWBC for each of the five time intervals specified above.

C. Use PROC PHREG to fit the extended Cox model described in Part-A and to estimate the hazard ratios described in Part-B.

D. Use PROC PHREG to fit a stratified Cox model that stratifies on SEX but keeps RX and LOGWBC in the model.

E. Compare the results from the models that you fit in Part-C and Part-D regarding the hazard ratio for the effect of RX? Is there any way to determine which set of results is more appropriate? Explain.

Question # 4: **12 points**

The data for this question contains survival times of 65 multiple myeloma patients and includes the following variables:

SURVTIME = Survival time (in months) from time of diagnosis
STATUS = Survival status (0 = alive, 1 = dead)
PLATELTS = Platelets at diagnosis (0 = abnormal, 1 = normal)
AGE = Age at diagnosis (years)
SEX = Sex (0 = female, 1 = male)
PLTAGE = Platelets by age interaction term (PLATELTS x AGE)
PLTSEX = Platelets by sex interaction term (PLATELTS x SEX)

Suppose that PROC PHREG was used to fit several different Cox models to this dataset (See next page).
Use the results from these models to answer the questions below.

A. State the form of the model being fit, in terms of the hazard function, for each of the five models.

B. For each of the five models, give an expression for the hazard ratio for the effect of the platelet variable (adjusted for age and sex, if included in the model).

C. Using your answer to Part-B, compute the estimated hazard ratio comparing a 40 year old male with normal platelets to a 40 year old male with abnormal platelets in each of the five models.

D. Based on the results displayed in Model 1, do you think that there are any significant interactions? Provide justification for your answer.

E. Considering Models 2-5, do you think that age and sex need to be controlled for as confounders?

F. Which of the five models would you choose to report? Why?

G. Based on the model selected in (f), write a short paragraph summarizing the relationship between the platelet variable and survival.

Model 1:

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
PLTLTS	1	0.470	2.854	0.027	0.869	1.600	0.006	429.689
AGE	1	0.000	0.037	0.000	0.998	1.000	0.930	1.075
SEX	1	0.183	0.725	0.064	0.801	1.200	0.290	4.969
PLTAGE	1	-0.008	0.041	0.036	0.850	0.992	0.915	1.075
PLTSEX	1	-0.503	0.804	0.391	0.532	0.605	0.125	2.924

Model 2:

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
PLTLTS	1	-0.725	0.401	3.260	0.071	0.484	0.221	1.063
AGE	1	-0.005	0.016	0.110	0.740	0.995	0.965	1.026
SEX	1	-0.221	0.311	0.503	0.478	0.802	0.436	1.476

Model 3:

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
PLTLTS	1	-0.706	0.401	3.106	0.078	0.493	0.225	1.083
AGE	1	-0.003	0.015	0.047	0.828	0.997	0.967	1.027

Model 4:

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
PLTLTS	1	-0.705	0.397	3.148	0.076	0.494	0.227	1.075
SEX	1	-0.204	0.307	0.442	0.506	0.815	0.447	1.489

Model 5:

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
PLTLTS	1	-0.694	0.397	3.065	0.080	0.500	0.230	1.088

Question # 5: **20 points**

The datasets discussed in the lecture on 'Time-dependent covariates' have been uploaded on Blackboard in the DATASETS folder. Download these datasets and try out the SAS codes we discussed in class. In addition, read Page #154 – 172 of the SAS book Survival Analysis Using SAS – by Paul D. Allison and practice using some alternate SAS codes provided by the author.

A. Confirm that you have understood how to code both ways: [i] inside PROC PHREG and [ii] using Counting Process input – while dealing with time-dependent covariates. Also confirm that you have understood the interpretation of the results (You don't have to turn in your work here, just mention that you completed this task).

B. For the Recidivism study that looks at effect of employment on time to arrest, Page #165 shows the SAS code for modeling "Cumulative proportion of weeks worked" as a time-dependent covariate. This code uses programming inside a DATA step before programming inside PROC PHREG.

   i. Run this code in SAS and interpret the results.
   ii. Then write your own code to program fully inside PROC PHREG. Run your code and match the results with those obtained from {i} above.

C. For the Recidivism study, researchers wish to assess if "employment stability" is associated with the time to arrest. Following two ways to capture employment stability have been proposed:

i. Using "Number of Switches" (A switch occurs when the employment status in a given week is different from the employment status in the preceding week) up to a given time.

ii. Using "Number of Negative Switches" (A negative switch occurs when a person who had employment in the previous week loses it in the current week i.e. the number of times a job was lost) up to a given time.

Note that just like in Part-B, you have to adjust for all other covariates.
Write SAS code to help researchers implement {i} and {ii} above. Interpret the results from your model.

Question # 6: BONUS **5 points**

Repeat Q#5 Part-C {ii} above treating "Number of Negative Switches" as a categorical variable with three categories: # times Job lost = 0, # times Job lost = 1, # times Job lost >=2

GOOD LUCK ☺☺