STA 312f23 Assignment Eleven[1]

The paper and pencil part of this assignment is not to be handed in. It is practice for Quiz 11 on December 1st. The R parts may be handed in as part of the quiz. **Bring hard copy of your printouts for Questions 2 and 4 to the quiz**. Separate printouts might be a good idea. Do not write anything on your printouts in advance except possibly your name and student number. *Answers to the "plain language" questions are specifically prohibited.* Do not write the answers to the plain language questions, or type them, or otherwise cause them to appear on your printouts.

1. In a study of whether vitamin C can help protect against the common cold, volunteer subjects were randomly assigned to either an observation only condition (no pill), a placebo condition (sugar pill), a 100 mg pill, a 500 mg pill or a 2,000 mg pill. The pills all had the same size and appearance. Time until the first reported cold was recorded. Subjects were followed for 6 months. At that point, the data for any subject who had not caught a cold was censored. In addition to experimental condition, age was recorded for each subject. Naturally, a lot more data about the subjects would be recorded in a real study.

   (a) These data will be analyzed using proportional hazards regression. Write the hazard function, including the baseline hazard. Denote age by x, and put it first. There are no interactions. You do not have to say how your dummy variables are defined. You will do that in the next part.

   (b) In the table below, make columns showing how your dummy variables are defined. Make placebo the reference category. In the last column, write the hazard function. If *symbols* for your dummy variables appear in the last column, the answer is wrong.

		Hazard Function
Observation Only		
Placebo		
100 mg		
500 mg		
2,000 mg		

(c) You want to test whether, controlling for age, experimental condition has any effect on the risk (technically, the hazard) of getting a cold. Experimental condition includes Observation Only.

i. Write the null hypothesis in scalar form, using the notation of your answer to Question 1a.

ii. Write the null hypothesis in matrix form as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$.

(d) Placebo effects are interesting. It is well documented that sometimes, people can get better by taking a pill that is medically inert, just because they *believe* it will make them better. It is particularly likely in this study, because we have data only on time to the first *reported* cold. You want to test for a placebo effect.

i. Write the null hypothesis in scalar form, using the notation of your answer to Question 1a.

ii. Write the null hypothesis in matrix form as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$.

(e) You want to test whether, controlling for age, dosage level has any effect on the risk of getting a cold. This test includes the placebo, which is a dosage level of zero, but excludes Observation Only.

i. Write the null hypothesis in scalar form, using the notation of your answer to Question 1a.

ii. Write the null hypothesis in matrix form as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$.

(f) If that last test is statistically significant, you definitely want to follow up with all pairwise comparisons of dosage levels, to see where the overall difference comes from. How many pairwise comparisons are there? The answer is a number.

(g) One of the pairwise comparisons is between the placebo and 2,000 mg.

i. Write the null hypothesis in scalar form, using the notation of your answer to Question 1a.

ii. Write the null hypothesis in matrix form as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$.

(h) Another pairwise comparison is between 100 mg and 500 mg.

i. Write the null hypothesis in scalar form, using the notation of your answer to Question 1a.

ii. Write the null hypothesis in matrix form as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$.

2. The `channing` data set in the `KMsurv` package has data on death in nursing homes as a function of age and sex. You will need to install the `KMsurv` package. Information about the data is provided by `help channing`. For some reason you need `data(channing)` before you can use it.

   Please make gender a binary variable with 1=Female and 0=Male. Also, create a centered version of age, by subtracting off the sample mean for the entire sample. This variable should be in years, not months. There are two age variables in the data file: `age` (age when death or censoring occurred), and `ageentry` (age at entry into the nursing home). Which one should you use? Well, think of the denominator of a term in the partial likelihood. You are summing over the data from individuals at risk at time (i) — that is, from individuals who have not died or been censored yet. Does it make sense for them to have covariate values for events that happened in the future? Based on this, pick the right age variable. From now on unless otherwise specified, "age" refers to this variable, centered.

   (a) In any data analysis, you should start with some basic descriptive statistics to see what the data are like. Please answer these questions. The answers are numbers. A `summary` of the data frame will help.

      i. What was the median age at which participants entered the nursing home, in years?
      ii. What was the age of the youngest person to enter the nursing home, in years?
      iii. What was the age of the oldest person to enter the nursing home, in years?
      iv. What proportion of the nursing home residents were women?
      v. What was the longest length of stay at the nursing home, in years?
      vi. What proportion of the observations were censored?

   (b) Fit a proportional hazards model with just age and gender as explanatory variables. The `time` variable represents not how long they survived in their whole lives, but how long they lasted at the nursing home. Let's use that as the response variable.

      i. True or False: The baseline hazard function is the hazard function for a woman of average age. Show a little work and answer True or False.
      ii. Interpret both Z-tests in plain, non-statistical language. This is one sentence each.
      iii. Correcting for age, the estimated hazard of death is _____ as great for women.
      iv. Give a 95% confidence interval for that last answer. Allowing for age, we estimate the hazard to be between _____ and _____ as great for women.
      v. Controlling for gender, if age at entry is increased by 10 years, we estimate the hazard of death to be multiplied by _____.
      vi. Give a 95% confidence interval for that last number.

vii. Suppose you had left gender as 1-2 and you left age in months, uncentered. Would this affect the test statistics and the conclusions? Run a quick example to find out. I used the original `channing` data set.

viii. Estimate the median survival times for men of average age and women of average age. Oops! I can't get an estimate for women.

ix. Make a well-labelled plot showing the estimated survival functions for a man and a woman of average age. There should be two different line types, like maybe solid and dashed. Print the graph and bring it to the quiz. From the picture, can you see why there was no estimated median for women?

x. The following is just a comment, not a question. A crucial assumption of the censored data model is that the censoring mechanism is independent of the outcome. However, lots of nursing home residents leave the home to go to hospital, and they die there rather than at the home. In fact, they are censored *because* the staff thought they were probably about to die. To me, this helps explain the high estimated probabilities of lasting a long time that we see in these data.

3. It is very natural and tempting to use age as a time-varying covariate. This question tries to illustrate why it won't work. Consider a model with an indicator for gender, and also age. Age $= x + t$, where x is age at entry to the study. The partial likelihood can be written

$$\begin{aligned} \mathrm{PL}(\boldsymbol{\beta}) &= \prod_{i=1}^{D} \left(\frac{e^{\mathbf{x}(t_{(i)})^\top \boldsymbol{\beta}}}{\sum_{j \in R_{(i)}} e^{\mathbf{x}(t_{(i),j})^\top \boldsymbol{\beta}}} \right) \\ &= \prod_{i=1}^{D} \left(\frac{e^{\beta_1 s_{(i)} + \beta_2 (x_{(i)} + t_{(i)})}}{\sum_{j \in R_{(i)}} e^{\beta_1 s_{(i),j} + \beta_2 (x_{(i),j} + t_{(i)})}} \right). \end{aligned}$$

Note that the event happens to individual (i) at time $t_{(i)}$, and the explanatory variables in the risk set are all assessed at exactly that same instant $t_{(i)}$. What happens?

4. If you spend time on the right social media sites, you will have heard of Area 51, a restricted region in the Nevada desert. It is widely believed that if you go hiking in Area 51, you have a good chance of being kidnapped by space aliens. There are indications that whether you are wearing a hat matters. To test this idea and in the interests of transparency, volunteers (there are plenty) went walking in the desert under controlled conditions.

   Each day for up to 30 days, the volunteer was dropped off by helicopter at a random location in Area 51. The volunteer was either wearing a hat or not, determined by a coin toss at the beginning of each day. Area 51 is out of ordinary cell phone range, but the U. S. military has a cell phone tower there, enabling the experimenters to stay in contact with the volunteers, and to determine their exact location.

   Kidnapping has a distinct signature. The volunteer's cell phone signal is suddenly interrupted. A helicopter is dispatched immediately to their last location and a search is initiated, but the volunteer is never found.

   If a volunteer is not kidnapped on a particular day, then after exactly eight hours, the helicopter picks up the volunteer for transport back to the base. If the volunteer is not kidnapped within 30 days, the observation is censored. Censoring can occur earlier if the volunteer withdraws from the study, or suffers a medical emergency.

   The file https://www.utstat.toronto.edu/brunner/data/legal/area51.data.txt contains a subset of the data from 200 volunteers, in a stop-start format. Variables include age, sex, an indicator for wearing a hat (which varies over time), event times, and a binary variable called taken, which equals one if a kidnapping occurred, and zero if there was no kidnapping. Event times are in minutes, with 8*60=480 minutes in an eight hour day. The clock stops at the end of each eight hour day, and re-starts when the volunteer is dropped off in the desert the next morning. Thanks to General Buck Turgidson for permission to use these data.

   Using proportional hazards regression, fit a model in which the explanatory variables are age, sex, and wearing a hat. The main objective is to determine whether wearing a hat has any effect on the risk of being kidnapped by aliens, but age and sex are interesting too. Be able to answer the usual questions about hazard ratios, and of course be able to state your conclusions in plain, non-statistical language.

Please bring **both** printouts to the quiz. Your printout should show *all* R input and output,and *only* R input and output. Do not write anything on your printouts in advance except your name and student number. The rule is that you may not put anything on your printout that you could not have known before seeing the results. So question numbers are okay. You may even copy-paste the entire questions (for the computer parts) into comment statements if you wish. But results, conclusions and interpretation are not allowed.