STA 312f23 Assignment Ten[1]

The paper and pencil part of this assignment is not to be handed in. It is practice for Quiz 10 on November 24th. The R parts may be handed in as part of the quiz. **Bring hard copy of your printout for Questions 5 and 6 to the quiz**. 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 them, or type them, or otherwise cause them to appear on your printouts.

1. Our main concrete example of a proportional hazards regression model is Weibull regression.

   (a) What is the baseline hazard function for Weibull regression? Assume e^{β_0} is part of the baseline hazard function.

   (b) Suppose that the Weibull regression model is the true model for a set of data. When we fit a proportional hazards regression model by maximum partial likelihood and estimate β_1, what function of the Weibull regression model parameters are we estimating?

2. Prove $S(t) = e^{-H(t)}$, where $H(t) = \int_0^t h(y)\,dy$. This is a general statement, not just for the proportional hazards model.

3. For the proportional hazards model, again assume that e^{β_0} is part of the baseline hazard function. We will always do this from now on. Prove that for the proportional hazards model, $S(t) = S_0(t)^{\exp\{\mathbf{x}^\top\boldsymbol{\beta}\}}$.

4. A sample of lung cancer patients are classified according to their type of cancer: squamous, small cell, adenocarcinoma, and large cell. We also have age and physician's rating of how far the disease has progressed on a scale from 1-10, which we will call "severity." Small cell lung cancer is found exclusively in smokers, ex-smokers, and people who have worked in the asbestos industry. *For this entire question, assume a proportional hazards regression model.*

   (a) Write the hazard function, the length of time between diagnosis and death (call it survival time) by t. Denote age by x_1 and disease severity by x_2. There should be *no interactions* in the model, in case you know what that is. You do not need to say how the dummy variables are defined. You will do that in the next part. Complete the equation below.

   $h(t) =$

(b) In the table below, make columns showing how your dummy variables are defined. In the last column, write the hazard function $h(t)$ given a particular vector of explanatory variable values $\mathbf{x}$, using the notation of your model from Question 4a above. If *symbols* for your dummy variables appear in the last column, the answer is wrong.

		$h(t)$
Squamous		
Small Cell		
Adeno		
Large Cell		

(c) In the notation of your model, what is the hazard function for a 45-year-old patient with adenocarcinoma and a disease severity of 6?

(d) For a patient with small-cell lung cancer, the hazard of death is _____ times as great as the hazard for a patient with large-cell lung cancer. Answer in terms of the Greek letters from your model. Do age and disease severity affect the answer (in this model)? Does time t affect the answer?

(e) For a 47-year-old patient with squamous lung cancer and a disease severity of 3, the chances of death are _____ times as great as the chances for a 47-year-old with adenocarcinoma and a disease severity of 3. Answer in terms of the Greek letters from your model.

(f) You want to know whether, controlling for age and disease severity, type of lung cancer has any effect on the risk of death. What is the null hypothesis? Answer in terms of the Greek letters from your model.

(g) That last question could be answered with either a large-sample likelihood ratio test, or a Wald test.

i. Suppose you decided on a likelihood ratio test. Write hazard function for the restricted model.

$h(t) =$

ii. Suppose you decided on a Wald test. Write the null hypothesis $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{0}$ in matrix form.

(h) You want to know whether, allowing for type of cancer and disease severity, the patient's age has any connection to risk of death. What is the null hypothesis? Answer in terms of the Greek letters from your model.

(i) You want to know whether, controlling for age and disease severity, the hazard is different for patients with large-cell or small-cell cancer. What is the null hypothesis? Answer in terms of the Greek letters from your model.

(j) You want to know whether, controlling for age and disease severity, risk is different for patients with squamous lung cancer or adenocarcinoma. What is the null hypothesis? Answer in terms of the Greek letters from your model.

5. The classic data set `veteran` is available as part of the `survival` package. Type `help(veteran)` for details. Look at `contrasts(veteran$celltype)` to see how the dummy variables for cell type are set up. It's not what you might expect.

(a) Based on a preliminary analysis (one that you don't have to do), I request that you fit a proportional hazards regression model with just experimental treatment, cell type and Karnofsky score. Based on this model, carry out significance tests to answer the following questions. Be able to state H_0, give the value of the test statistic (Z or chi-squared) and the p-value. Be able to state your conclusions, if any, in plain, non-statistical language. Guidelines for the plain language statements are

- Be guided by the 0.05 significance level, but never mention it. If you do, you get a zero even if what you say is correct.
- Any use of statistical vocabulary such as p-value, null hypothesis, significance etc. will get you a zero. Instead of saying "controlling for," say "allowing for," or "correcting for," or "taking into account." The phrase "controlling for" will not get you a zero, but please avoid it when talking to non-statisticians.
- If a directional conclusion is posible, make it. Don't say "Survival time was related to sex." Say "Women tended to live longer."
- If a test is not significant, do *not* say there was no effect, or no difference. Avoid accepting the null hypothesis, or implying that you accept it. Say "There was no evidence that surgery was related survival time," or "These results do not provide evidence of a connection between marital status and time required to graduate," or something like that.
- For any explanatory variable that was *not* randomly assigned, avoid language that suggests influence, or causal connection. Say "Patients with a health club memberships were at less risk for heart attack," not "Exercise lowered the risk of heart attacks."

Now here are the questions.

i. Controlling for cell type and Karnofsky score, does treatment appear to affect survival time?

ii. Allowing for experimental treatment and cell type, does Karnofsky score help predict survival? In spite of the word "predict," you are beng asked for a significance test.

iii. Correcting for experimental treatment and Karnofsky score, do patents with different types of cancer (cell type) differ in their hazard of dying? Do a partial likelihood ratio test.

iv. Follow up the last question by carrying out tests for all pairwise comparisons of cancer types, controlling for the other variables. Some of the comparisons you want are z-tests on the `summary` output. Use Wald tests for the other comparisons. Directional conclusions are possible for all the tests that are statistically significant, including the Wald tests.

6. This question uses the same old `cancer` data set you have been analyzing in the past two assignments. Even though you may be getting tired of it, there is an interesting technical question we have not explored. *Please print the output for this question on a separate sheet.*

   (a) Fit a proportional hazards model using the same explanatory variables you did for Weibull regression and log-normal regression: `sex` and `ph.ecog`. Be able to state the conclusions in plain, non-statistical language. See Question 5a for guidelines.

   (b) Now do a table of `ph.ecog`. Also do `summary`. You can see that even though it's technically a 6-point scale, in practice the physicians are using just a few categories. It makes you wonder whether we should be treating `ph.ecog` as quantitative or categorical.

   (c) Fit a model with `sex` and `ph.ecog`, in which `ph.ecog` is represented by dummy variables. Before you do this, make `ph.ecog = 3` into `NA`, since there's only one patient.

   (d) If `ph.ecog` is quantitative, the proportional hazards model says the hazard is multiplied by e^{β_2} when we increase by one unit (whether it's from 0 to 1 or from 1 to 2). Express this as a null hypothesis about the parameters of your model from Question 6c.

   (e) Test the null hypothesis with a Wald test. What do you conclude? (This is *not* a plain language question.) Does it seem okay to treat `ph.ecog` as quantitative?

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. In particular, answers to the "plain language" questions must not appear on your printout.