STA 312f23 Assignment Eight[1]

The paper and pencil part of this assignment is not to be handed in. It is practice for Quiz 8 on March 11th. The R part may be handed in as part of the quiz. **Bring hard copy of your printout to the quiz**. Do not write anything on your printout in advance except possibly your name and student number.

1. Consider the multiplicative regression model for the failure time: $t = e^{\beta_0+\beta_1 x} \times \epsilon$, where β_0 and β_1 are unknown constants (parameters), x is a known, observed constant, and $\epsilon \sim \exp(1)$.
   (a) Derive the probability density function of t. Do it directly, not the way it was done in the lecture slides.
   (b) Using the formula sheet, write down the
      i. Expected value of t.
      ii. Median of t.
      iii. Survival function of t.
   (c) Give the hazard function of t. Show some work.

2. Let $\epsilon \sim \exp(1)$.
   (a) Derive the density of $W = \epsilon^\sigma$, where $\sigma > 0$.
   (b) That density is Weibull. What are the parameters α and λ?

3. Let W have a Weibull distribution with parameters α and λ, and let $T = c\,W$, where c is a positive constant. Show that the distribution of T is Weibull, and give the parameters.

4. Let the failure time $t_i = \exp\{\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1}\} \cdot \epsilon_i^\sigma$, where again, $\epsilon \sim \exp(1)$.
   (a) Based on your answer to the preceding questions, what is the distribution of t_i? Just write down the answer.
   (b) Show that if $x_{i,k}$ is increased by c units, t_i is multiplied by $e^{c\beta_k}$.

5. For the model of Question 4, give the following. Don't do more calculation than you have to.
   (a) Expected value of t.
   (b) Median of t.
   (c) Survival function of t.
   (d) Hazard function $h(t)$. Show your work, and then check the formula sheet.
   (e) If $x_{i,k}$ is increased by one unit, the hazard function is multiplied by ______. Show some work if you need to. The answer is *not* e^{β_k}.

6. Show that the Weibull model of Question 4 has proportional hazards. That is, consider the hazard functions of two individuals with different $\mathbf{x}$ vectors. Show that the ratio of their hazard functions does not depend on t. This means that the two hazard functions are always in the same proportion at every point in time.

[1] This assignment was prepared by Jerry Brunner, Department of Mathematical and Computational Sciences, University of Toronto. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/brunner/oldclass/312f23

7. 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.

   (a) Write a (multiplicative) Weibull regression equation, denoting the length of time between diagnosis and death (call it survival time) for patient i by t_i. Denote age by $x_{i,1}$ and disease severity by $x_{i,2}$. 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.

   $t_i =$

   (b) In the table below, make columns showing how your dummy variables are defined. Make small cell the reference category. In the last column, write the expected survival time, using the notation of the model of Question 4. If *symbols* for your dummy variables appear in the last column, the answer is wrong.

		Expected Survival Time
Squamous		
Small Cell		
Adeno		
Large Cell		

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

   (d) You want to produce a large-sample confidence interval for expected survival time, for a 50-year-old patient with adenocarcinoma and a disease severity of 2. You need to use the delta method.

      i. What is the parameter vector $\boldsymbol{\theta}$? Give a general answer for your model.

      ii. What is $\dot{g}(\boldsymbol{\theta})$ for this particular example?

   (e) For 47-year-old patients with small cell lung cancer and a disease severity of 3, the median survival time is __________ times as great as the median survival time for 47-year-old patients with large cell lung cancer and a disease severity of 3. Answer in terms of the Greek letters from your model.

(f) For a patient with squamous lung cancer, expected survival time is __________ times as great as the expected survival time for a patient with adenocarcinoma, given the same age and disease severity. Answer in terms of the Greek letters from your model. According to the model (not the same thing as reality), do the exact values of age and disease severity affect the answer?

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

(h) 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 the multiplicative Weibull regression equation for the restricted model.

   $t_i =$

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

(i) You want to know whether, allowing for type of cancer and disease severity, the patient's age has any connection to life expectancy. 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, median survival time 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.

(k) You want to know whether, controlling for age and disease severity, median survival time 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.

8. The `survival` package has a built-in data set on patients with advanced lung cancer. Type `help(cancer)` for details.

   (a) An important question is whether self-ratings by the patients are useful predictors of survival. Ignoring all other variables for the present, fit a Weibull regression model with just one explanatory variable: Karnofsky performance score as rated by patient.

      i. Give the value of the test statistic, and also the p-value. Do you reject the null hypothesis at the $\alpha = 0.05$ significance level? State the conclusion in plain, non-statistical language.

      ii. Here is a technical question. Remember that the exponential distribution is a special case of the Weibull. The output of `summary` includes a test that tells you whether a Weibull regression model fits significantly better than an exponential regression model. Briefly explain. Give the value of the test statistic, and also the p-value. Do you reject the null hypothesis at the $\alpha = 0.05$ significance level? What do you conclude?

(b) Now fit a Weibull regression model with all the available explanatory variables, excluding institution. Controlling for all other explanatory variables, is the Karnofsky performance score as rated by the patient related to survival time? Give the value of the test statistic, and also the p-value. Do you reject the null hypothesis at the $\alpha = 0.05$ significance level? State the conclusion in plain, non-statistical language.

(c) You will now observe something that has led to many incorrect conclusions based on likelihood ratio tests. Without dropping any variables at this time, test the diet and weight loss variables in a single test, controlling for all the other explanatory variables. Do it two ways, with a likelihood ratio test and a Wald test. Guided by the $\alpha = 0.05$ significance level, what do you conclude in each case? Do the results agree? Which one is more compatible with the results of the Z-tests?

(d) If you have not done it already, look at `summary` for your reduced model above. Do you see the sample size? Compare this to the sample size for the full model.

The default behaviour of `survreg` is to omit cases with any missing values. Look at the output of `summary(cancer)`, something you should have looked at carefully *before* starting the analyses. Do you see the missing values for `meal.cal` and `wt.loss`? Recalling that the likelihood ratio test statistic G^2 is the difference between two -2 log likelihoods, explain how the likelihood ratio test is affected by missing values in the variables that are omitted from the restricted model.

(e) The solution is to fit both the full model and the restricted model using a data set that has *no missing values for any of the variables in the full model.* Create a data frame with this property. Yes, specifically a data *frame*. See `help(na.omit)`. Please note that you do not want to omit the one case that has institution missing, but no other missing data. My data frame has 168 rows.

(f) Based on this new data frame with no missing values, fit the full and restricted models, and test the difference between them with a likelihood ratio test. My test statistic value is $G^2 = 3.279398$. Are these results closer to the Wald test?

(g) Now drop `age`, `pat.karno`, `meal.cal` and `wt.loss`, obtaining a smaller model that we hope is cleaner and better for prediction. Larger sample size is good here. Don't throw any data away if you can help it.

Looking at the `summary` output for this model, one is forced to wonder whether it's necessary to make busy doctors fill out two questionnaires instead of just one, especially since the some of the questions are likely very similar.

Based on this consideration, drop another variable. We now have a model with just two explanatory variables. Fit that model and look at summary.

i. Controlling for physician's rating of how poorly the patient is doing, is median survival time different for males and female patients? Give the null hypothesis in symbols, the value of the test statistic and the p-value (numbers), and state the conclusion in plain, non-statistical language. Use the word "average" instead of median. After all, your conclusion also applies to the expected value.

ii. Allowing for patient's gender, is physician's rating informative about survival time? Give the null hypothesis in symbols, the value of the test statistic and the p-value (numbers), and state the conclusion in plain, non-statistical language. Always draw directional conclusions when possible.

iii. The default output of summary includes a test that will allow you to decide whether the hazard function is increasing or decreasing, without actually plotting it. Do the necessary paper-and-pencil calculation (one line). Then, state the null hypothesis in symbols, give the value of the test statistic and the p-value (numbers from the printout), and state the conclusion in plain, non-statistical language. Do not use the term "hazard function."

iv. Estimate the median survival time for female patients with an `ecog` rating of 1. Include a 95% confidence interval.

v. This analysis could continue. Look at `table(ph.ecog)`. What is the next thing you would do with the data?

Please bring your printout 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 except your name and student number.