STAT 437 - Assignment 3
Due: Friday, March 18 on Crowdmark
Resubmit: Friday, April 1 on Crowdmark

This assignment covers the materials contained in **Lecture 029** through to **Lecture 045**. Reminder that you are permitted to discuss to these problems with classmates, but every student must submit their own solutions which are their own work (including any code, figures, etc.). **Please indicate any students that you discussed solutions with on your submission**. Please ensure that your submissions on Crowdmark are legible, and separated based on the problems included at the submission link. Submissions can be handwritten or typeset.

Part 1: True or False (20 Marks)

For each of the following problems indicate whether the statement is true or false, and give a short justification for your answer. Correct answers without justification will receive only partial credit.

Problem 1. Suppose you are presented with the following life table results from a discrete time study.

Time	Number at Risk	Number of Events
1	250	22
2	228	13
3	215	5
4	210	26
5	184	11
6	A	B

True or False: From the given information we can conclude that $A = 173$.

Problem 2. A study was conducted which measured disease-free survival for patients following treatment, measured in years. The study commenced with 337 individuals, and information regarding the number of observed events and individuals being censored during the first four years are included in the following table.

Time	Number Censored	Number of Events
1	7	18
2	15	41
3	48	46
4	24	36

True or False: We can conclude that $\widehat{h}(3) - \widehat{h}(2) = \frac{241}{4992}$.

Problem 3. Suppose that T is a continuous, time-to-event, random variable. Moreover, suppose you are given the hazard function of T as $h(t)$. **True or False: This would be sufficient information to compute $E[T]$.**

Problem 4. A proportional odds model is fit to some data, which results in the following estimating results:

$$\text{logit}\left\{\widehat{h}(j)\right\} = 0.51D_1 + 0.99D_2 + 0.48D_3 - 0.13D_4.$$

Here, $D_j = I(t = j)$. **True or False: Based on the estimated model, we can find that $\widehat{S}(4) = 0.02068788$.**

Problem 5. Suppose you are conducting a study to determine the age at which individuals begin to smoke, and so you begin to survey individuals aged 18+, and then continue following them for the duration of your study. You note that the data you observe may be subject to both left-censoring and right-censoring. Upon completion of the study period, you compute that the sample mean age for beginning to smoke is 23.2. **True or False: This is an underestimate of the true value, since censored data push the mean estimate to be lower than it otherwise would be.**

Problem 6. Statistics Canada reports several key statistics regarding post-secondary students' educational attainment, including average time to graduation[1]. The above page mentions that

> The average time to graduation indicator represents the average number of elapsed academic years new students took to complete the credential in a given educational qualification. [...] students who did not complete the educational qualification within the period of observation were excluded from the average time to graduation calculation.

In other words, students who have not graduated at the end of the observation period are excluded from the calculation of average time to graduation. The observation window is defined to be 1.5 times the typical duration of a program. Assume that no students are lost to follow-up in this study (that is, every student is observed until they either graduate or drop out of post-secondary education). **True or False: It is accurate to describe this as the average time to graduation for post-secondary students in Canada.**

Problem 7. Suppose that you use a proportional odds model where the variate of interest is treatment group. You find that, for the individuals receiving the experimental treatment, $\widehat{h}_E(1) = 0.958$, $\widehat{h}_E(2) = 0.685$, and $\widehat{h}_E(3) = 0.237$. For those who received the placebo treatment, you compute $\widehat{h}_P(2) = 0.312$. **True or False: From this information you know that the proportional odds model would estimate that $\widehat{h}_P(1) = 0.4363445$ and $\widehat{h}_P(3) = 0.1079474$?**

Problem 8. Suppose that you fit a proportional hazards model and find that the coefficient on a treatment indicator is equal to -2.94. **True or False: This means that the hazard ratio for an individual who received the treatment, compared to one who did not receive the treatment, is equal to -2.94.**

[1] https://www150.statcan.gc.ca/n1/pub/37-20-0001/372000012021004-eng.htm

Problem 9. Suppose that you get the following output from a model fit using servreg in R.

```
## Call:
## survreg(formula = Surv(time, status) ~ age + factor(ses),
##         data = birth, dist = "lognormal")
##                      Value       Std. Error
## (Intercept)           0.2305      0.07599
## age                   0.0187      0.00276
## factor(ses)lower      0.0512      0.03001
## factor(ses)unknown    0.0824      0.03712
## factor(ses)upper     -0.0188      0.09099
## Log(scale)           -0.5962      0.01756
```

True or False: The implied error distribution from this model is normally distributed with $\sigma^2 = 0.35545444$.

Problem 10. Suppose that you fit both a proportional hazards model and an AFT model to the same dataset. Some of the resulting coefficients are summarized in the following table.

Model	Parameter	Value
Proportional Hazards	Coefficient for Age	0.41
Proportional Hazards	Log(Scale)	0.246
Accelerated Failure Time	Coefficient for Age	-0.89

True or False: It is possible that you used a Weibull distribution for both models.

Part 2: Conceptual Question (10 Marks)

For the following questions, provide your answers with justification and clear communication. The answers do not need to be long, but correct responses without complete justification will receive only partial credit.

Working as a data scientist, you are tasked to review and help to implement an analysis of your company's sales data. In particular, you have records of the different sales interactions that prospective customers had over a period of time, as well as relevant underlying demographic information. Each customer is recorded as having either made a purchase, having been lost to follow-up, or still active in the sales funnel at the time you are running your study. Sales in your company normally proceed over the course of several months. Your company is interested in using the data collected so far to get insight into the sales funnel, and try to establish best practices going forward. As a result, they want you to make any inferences and suggestions based on the data that you have available today.

Problem 11. (4 Marks) Indicate how the analysis of this data can be formulated in terms of survival analysis. Pay particular attention to specifying the necessary components for survival analysis, and ensuring that you indicate how each possible status of an individual corresponds to those which we can analyze in this framework.

Problem 12. (2 Marks) Your boss thinks that the sample mean time is a good way to capture customers' purchase behaviour. Is this the case? Explain using course concepts.

Problem 13. (2 Marks) Part of the sales process involves some customers receiving an in-person sales pitch. This is an incredibly costly process, which management is convinced helps the sales funnel, owing to a few high-profile customers who have made purchases shortly after receiving the pitch. Suppose that you fit an accelerated failure time model, which appears to be a good fit. The estimated coefficient for the in-person sales pitch is 0.70. Does this provide evidence that it is an effective strategy at increasing the speed of the sale?

Problem 14. (2 Marks) A coworker of yours, who has learned a little bit of survival analysis, suggests that you drop any individuals who are still in the sales funnel at the time of your analysis from consideration. How would this suggestion change your ability to draw conclusions?

Part 3: Theoretical or Applied Question (20 Marks)

Please **pick one** of the following two problems and solve it. **If you solve both** only the first problem will be looked at. The first problem corresponds to an **application** problem, while the second is a **theoretical question**. For the application question, please provide the code and relevant output (consider using a software like RMarkdown, or being highly selective with what output you copy to ensure your solution is legible). For the theoretical question, please include enough of your work to justify the steps you have taken.

PROBLEM 15. (Application) On the course website you will find a dataset called `stroke.csv`. These data consist of 518 acute stroke patients from a Copenhagen Stroke Study (Jørgensen et al., Stroke, 1996, 27(10): 1765–1769). The survival times are measured in years from the time of hospital admission. The data are recorded in a summary format, summarizing the number of individuals who had a stroke in each year (`d` in the data) out of all individuals at risk (`n` in the data), based on two separate grouping factors. First, individuals are indicated as either having had a previous stroke at the time of admission (`prevstroke` = 1 if yes, otherwise 0) and whether or not the individuals have a history of ischemic heart disease (`ihd` = 1 if yes, otherwise 0). Each row then consists of the number of at risk individuals and the number of events, for each year, within each group.

1. Estimate the survival probabilities for those who experienced a previous stroke, and those who did not, regardless of their IHD status. What is the estimated median survival time for both sets of patients?

2. Consider fitting a proportional odds discrete time model for these data which includes history for IHD, history of stroke, and the interaction. Does the proportional odds assumption seem valid? Discuss using a formal hypothesis test.

   Hint: Note that for grouped data, the GLM call in R is given by

   ```
   glm(cbind(d, n-d) ~ ..., ...)
   ```

   where d is the number of events and n is the total number at risk.

3. Is the model from (2) an appropriate model? Can it be simplified? Considering selection of the most appropriate model starting from the model you fit for (2).

4. Based on the fitted model in (3), discuss the impacts of disease and stroke history on the risk of death at a given year. Provide the corresponding estimates, 95% confidence intervals for these estimates, and a brief interpretation in words.

5. What is the median survival time for patients who have both a history of IHD and a history of previous strokes?

Problem 16. (Theoretical) Let T be a continuous random variable with $T \in [0, \infty)$. Take $f(t)$, $F(t)$, $S(t)$, $h(t)$, and $H(t)$ to be the density, cumulative density function, survival function, hazard function, and cumulative hazard function, respectively.

1. Demonstrate that $h(t)F(t) = h(t) - f(t)$.
2. Let $X = \min(T, C)$ where $C > 0$ is a pre-defined constant, and let $Y = H(X)$. Show that $E[Y] = F(C)$.

   Hint: the result from part (1) may be helpful.
3. Show the the cumulative distribution function of $Y = S(T)$ is uniformly distributed on $[0, 1]$.

   Hint: consider what we know about the monotonicity of $F(t)$.
4. Suppose that V has a uniform distribution on $[0, 1]$. Show that $X = -\log V$ has an exponential distribution with hazard rate equal to 1.