Exam 2
STA4712, Spring 2023
Introduction to Survival Analysis
Department of Statistics, University of Florida
Due 5:00PM on Tuesday, May 2nd, 2023

Please read the following instructions carefully before starting the exam.

- During the exam, you are allowed to consult your course textbook, course notes, or any other materials you have which may be useful. **Speaking with your classmates or anyone else (in-person, online, etc.), about the exam is strictly prohibited**. If it is deemed that you shared answers or spoke with others about the exam, you will receive a zero and will be reported to the Office for Student Conduct and Conflict Resolution. Do not put yourself in this situation.

- **The work you submit must be your own, original writing and code**. You should be especially skeptical of solutions provided by LLMs: `chatGPT-4`, for example, can give wildly incorrect answers to problems on this exam.

- Please make your exam as well-organized and easy-to-read as possible. If you perform mathematical derivations by hand and would like to include them in your exam, please scan them and paste the image of your work in the corresponding PDF. **You are to submit your completed exam as a single continuous PDF document to the course eLearning site. You must also submit an R markdown file used to create your pdf.** If you do math "by-hand" you may paste a scanned version of your handwritten math into your R markdown document using

  ```
  \includegraphics{handwrittenwork.pdf}
  ```

  where `handwrittenwork.pdf` is a pdf living in the same folder as your R markdown document. Note that only the relevant code should be included: if you include code which gives an error, for example, you may lose points.

- **Show as much work as possible.** Please give adequate explanation for everything you do, even if I do not explicitly ask for it. For instance, if I ask you for an estimated probability, please explain how you got that probability. I can't give you partial credit if I don't know what you tried, so it's in your best interest to be explicit about how you got your answers.

- To ask questions about the exam over email, use the subject line: **4712 Exam Question: [Your last name]**. I will respond to all emails within 12 hours, except on the 28th-29th.

- If no directions are explicitly given, you may use a $\alpha = 0.05$ significance level cutoff and construct 95% confidence intervals.

- This exam has four questions and is out of 54 points.

1. (19pts) When oncologists take a tumor sample, they are able to quantify the proportion of cells in the sample that are of a particular cell type (known as the "cell type composition"). For example, a particular tumor sample could have 20% T-cells[1] and 80% other cell types. In this problem, we will explore modeling the survival of a cancer patient as a function of their tumor sample T-cell proportion.

   The model we consider is the log-logistic accelerated failure time model

$$\log T = \beta_0 + \beta_1 C_T + \sigma\epsilon,$$

   where $C_T \in [0,1]$ is the proportion of cells which are T-cells; and (β_0, β_1) and σ are the unknown regression coefficients and scale parameter for the logistic error, respectively. We assume ϵ follows a logistic distribution with expected value zero and location parameter one, i.e., the density of ϵ is given by

$$f(x) = \frac{\exp(-x)}{\{1+\exp(-x)\}^2} \quad x \in (-\infty, \infty).$$

   Suppose I fit the model in R and obtain the following output.

```
> mod1 <- survreg(Surv(time, status) ~ CT, dist = "loglogistic")
> summary(mod1)

Call:
survreg(formula = Surv(time, status) ~ CT, dist = "loglogistic")
              Value Std. Error      z       p
(Intercept)  5.2638     0.0392 134.44 < 2e-16
CT           1.2404     0.3560   3.48 0.00049
Log(scale)  -2.0059     0.0756 -26.53 < 2e-16

Log logistic distribution
Loglik(model)= -607.3   Loglik(intercept only)= -613.2
    Chisq= 11.7 on 1 degrees of freedom, p= 0.00062
Number of Newton-Raphson Iterations: 7
n= 261

> vcov(mod1)
             (Intercept)            CT   Log(scale)
(Intercept)  0.001532878 -0.0118172164 0.0002267260
CT          -0.011817216  0.1267083935 0.0002232684
Log(scale)   0.000226726  0.0002232684 0.0057166497
```

   Answer the following questions based on the above output.

[1] T-cells play a central role in adaptive immune response and are characterized by the presence of a T-cell receptor on their surface.

(a) (3pts) Is the proportion of T-cells, C_T, significant? To receive full credit, write down the hypotheses you are testing, the distribution of your test statistic under the null hypothesis, provide the p-value, and state your conclusion in the context of the problem.

(b) (2pts) Provide a 95% confidence interval for the scale parameter σ.

(c) (3pts) Provide the maximum likelihood estimate of the survival function for T at time $t = 100$ for a patient with T-cell proportion $C_T = 0.20$ based on the output above.

(d) (3pts) Provide the maximum likelihood estimate of the 85th percentile of survival for a patient with T-cell proportion $C_T = 0.20$.

(e) (3pts) Provide a 95% confidence interval for the median survival time T for a patient with T-cell proportion $C_T = 0.20$.

(f) (3pts) What is the estimated hazard ratio for an individual with $C_T = 0.20$ versus an individual with $C_T = 0.05$? Note that this may be a function of time t.

(g) (2pts) Suppose we were not sure about the assumption of logistically distributed errors. Describe, in a few sentences, how I could check this assumption. (**Note.** You may assume that any model fitting strategy, parametric or otherwise, is available to you.)

2. (15pts) In this problem, we will model the time until heart failure of 299 patients. The predictors we will consider are

- `anaemia`: decrease of red blood cells or hemoglobin (0/1)
- `serum_creatinine`: level of serum creatinine in the blood (mg/dL)
- `ejection_fraction`: percentage of blood leaving the heart at each contraction (percentage)
- `high_blood_pressure`: if the patient has hypertension (0/1)
- `platelets`: platelets in the blood (kiloplatelets/mL)

For the first few parts of this problem, we will consider the Cox proportional hazard models

```
heart <- read.csv("https://ajmolstad.github.io/docs/heartData.csv", header=T)
heart$status <- heart$DEATH_EVENT
mod.q2 <- coxph(Surv(time, status) ~ anaemia + serum_creatinine + platelets +
                ejection_fraction + high_blood_pressure, data = heart)
```

(a) (2pts) If you look at the `summary` output from `mod.q2`, there is no intercept. In a few sentences, explain why.

(b) (3pts) Using a likelihood ratio test, test whether any of the coefficients from `mod.q2` are nonzero. For full credit, formally state the model you are testing, the null and alternative hypotheses, the distribution of the test statistic under the null, the p-value, and your conclusion.

(c) (2pts) Plot the estimated survival curves for a patient with `serum_creatinine = 1.1`, `ejection_fraction = 38.00`, and `platelets = 303500` for all four combinations of `anaemia` and `high_blood_pressure`. Describe what these plots tell you about the effects of high blood pressure and anaemia.

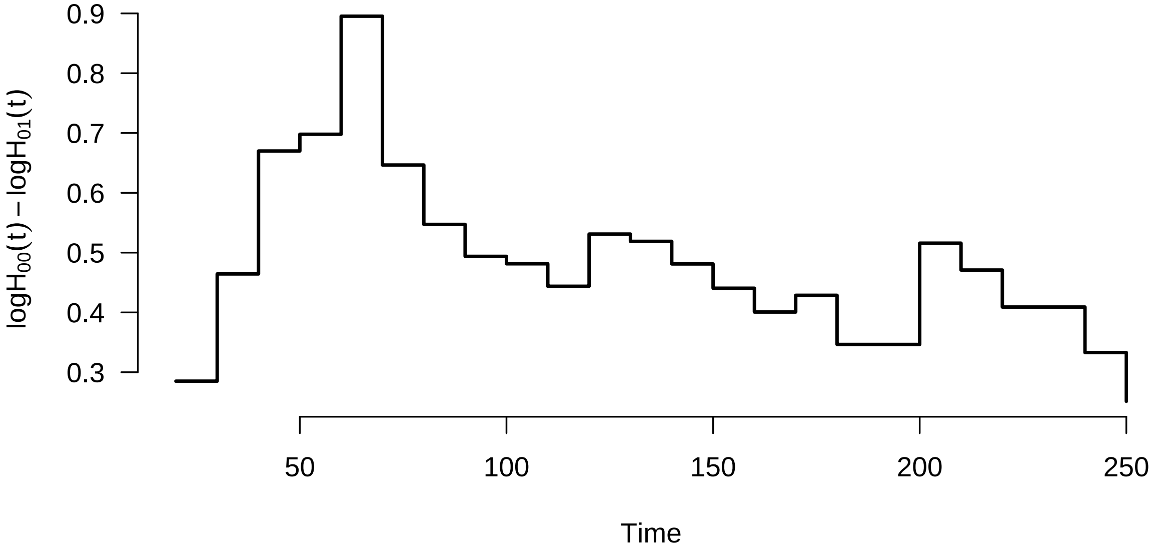
(d) (2pts) To check the model `mod.q2`, plot the standardized difference between the estimated cumulative hazard function and the cumulative hazard function of an exponential random variable. Based on the results, do you think the Cox PH model is appropriate for these data? Explain why you can conclude this based on your plot.

(e) (2pts) Of course, the Cox PH model assumes proportional hazards. For example, with all other covariates held fixed, the hazard with

`high_blood_pressure = 0` versus `high_blood_pressure = 1`

is constant across time. Let H_{00} and H_{01} be the baseline hazard function for the model with `high_blood_pressure = 0` and `high_blood_pressure = 1`, respectively. With the following code, we can make the figure below.

```
source("https://ajmolstad.github.io/docs/fun.R")
cumHaz <- function(sfit, t) {
  K <- length(sfit$strata)
  s <- c(0, cumsum(sfit$strata))
  H <- matrix(NA, nrow=length(t), ncol=K)
  for (i in 1:K) {
    ind1 <- (s[i] + 1):s[i + 1]
    tmp1 <- c(0,sfit$time[ind1])
    tmp2 <- c(0,sfit$cumhaz[ind1])
    H[,i] <- approxfun(tmp1, tmp2, method="constant")(t)
  }
  H
}
mod.q2.strat <- coxph(Surv(time, status) ~ anaemia + serum_creatinine +
    platelets + ejection_fraction + strata(high_blood_pressure), data = heart)
sfit <- survfit(mod.q2.strat)
Time <- seq(20, 250, by=10)
plot(Time, apply(log(cumHaz(sfit, Time)), 1, diff), type="s", bty="n", las=1,
    lwd=2,  xlab="Time", ylab=expression(log*H[0][0](t)-log*H[0][1](t)))
```



(e.i) In a few sentences, explain what the plot is displaying. Does it provide evidence that the Cox PH model assumptions are violated?

(e.ii) Under the assumption of proportional hazards, what is the maximum likelihood estimate of $\log H_{00}(t) - \log H_{01}(t)$?

(f) (2pts) Plot the martingale residuals (y-axis) versus the serum creatinine (x-axis). Color points according to whether the time was censored or represents the time of heart failure.

(g) (2pts) You should notice that only two individuals whose survival times we observed have martingale residuals less than -0.4 (`heart[c(49, 218),]`). Does our model over or underpredict these individuals' risk of heart failure? Can you provide any explanation or suggestion for why we over or under predicted for these two?

3. (9pts) Suppose we have the following times, indicators of censoring, and covariate triplets $\{(t_i, \delta_i, x_i)\}$,

$$\{(t_i, \delta_i, x_i)\}_{i=1}^{7} = (4,1,3), (5,0,1), (4,1,2), (1,0,2), (2,0,2), (2,1,3), (6,1,-1),$$

where $\delta_i = 1$ indicates an observed failure time and $\delta_i = 0$ a censoring time. You may download these data in R using

```
Q3dat <- readRDS(url("https://ajmolstad.github.io/docs/4712_Q3.RDS"))
```

We will assume a proportional hazards model for these data, i.e., the hazard function for a subject with predictor x_i is

$$h(t \mid x_i) = h_0(t)\exp(\beta x_i).$$

(a) (3pts) Because there are tied failure times in our data, the Cox partial likelihood is not well defined. Write down (i) the Efron approximation to the partial log-likelihood, and (ii) the Breslow approximation to the partial log-likelihood.

(b) (1pt) Are the maximum likelihood estimators of β under the two approximations to the partial log-likelihood equivalent?

(c) (3pts) Under the Breslow approximation to the partial likelihood, perform a likelihood ratio test of the hypotheses

$$H_0 : \beta = 1 \text{ versus } H_a : \beta \neq 1.$$

4. (11pts) In this problem, we will compare multiple parametric AFT models based on five predictors, $x_{i1}, \ldots, x_{i5}$, using the model

$$\log T_i = \beta_0 + \sum_{j=1}^{5} x_{ij}\beta_j + \sigma\epsilon_i.$$

We will consider three distributions for ϵ_i, which can be fit using the following code.

```
Q4dat <- readRDS(url("https://ajmolstad.github.io/docs/4712_Q4.RDS"))
q4.formula <- Surv(time, status) ~ X1 + X2 + X3 + X4 + X5
mod4.wei <- survreg(q4.formula, dist="weibull", data = Q4dat)
mod4.exp <- survreg(q4.formula, dist="exponential", data = Q4dat)
mod4.ln <- survreg(q4.formula, dist="lognormal", data = Q4dat)
```

(a) (2pts) Can I compare `mod4.wei` to `mod4.exp` using a likelihood ratio test? If so, perform this test and state your conclusion in the context of the problem. If not, explain why not.

(b) (2pts) Can I compare `mod4.wei` to `mod4.ln` using a likelihood ratio test? If so, perform this test and state your conclusion in the context of the problem. If not, explain why not.

(c) (2pts) Let γ_i denote the linear predictor for the ith subject from an AFT model (e.g., in the models above, $\gamma_i = \beta_0 + \sum_{j=1}^{5} x_{ij}\beta_j$). Suppose we have a pair of observations $(t_i, \delta_i, \gamma_i) = (100, 1, 0)$ and $(t_k, \delta_k, \gamma_k) = (150, 1, 1)$. Is this pair concordant, discordant, tied, or indeterminant? Explain your answer briefly.

(d) (2pts) Provide the AIC, BIC, and concordance in a table all three models. Which model is best according to AIC, BIC, and concordance? **Note.** Concordance for a model fit using `survreg` can be obtained using `concordance(mod4.a)`, for example.

(e) (3pts) (e.i) To see which model predicts best on a new dataset, perform 5-fold cross-validation, measuring concordance on the left-out fold for each model.[2] (e.ii) Report the average concordances for each of the three models. Which model is best? (e.iii) Do the models perform better than randomly guessing the order of survival times? How do you know?

[2] **Hint.** Be careful using `survConcordance` as it assumes the linear predictors you input come from a Cox model, not an AFT model. How can you remedy this?