STAT 486/886 (Winter 2022)

Assignment 4

The assignment is due on Apr. 10 (Sunday), 23:00 (time of Kingston Ontario Canada). Please submit to OnQ, the dropbox for this assignment.

Guidelines for Preparing Solutions:
For some problems in this course, complete code and output may be very long. Please only include important output and necessary results in the main text of your solutions.
Give descriptions and discussions for your important exploration and findings.
Put long, extra code and output in an **Appendix,** at the end of your assignment.
The Appendix will NOT be marked. Please include it as evidence of your independent work.
Prepare your assignment solutions so that it is easy for readers to follow, without having to search everywhere for your answers from lengthy code and output.
Students submitting identical solutions (multiple sentences, derivation steps or code copied among students or from other resources) will be investigated for violations of academic integrity.

1. Let T be the failure time of a patient with a single covariate z.
(a) Assume that T is from an accelerated failure time distribution. That is, $Y = \log T$ has a location-scale distribution, and

$$Y = \beta_0 + \beta_1 z + \sigma W, \tag{1}$$

for W from the standard location-scale distribution with $\mu = 0$ and $\sigma = 1$. Let T_0 be the failure time of a patient with baseline covariate value $z = 0$. Show that

$$S_T(t|z) = S_0(te^{-\beta_1 z}) \tag{2}$$

where $S_T()$ is the survivor function of T and $S_0()$ is the baseline survival function (i.e. the s.f. of T_0). How does the covariate z affect the survival distribution?
(b) Assume that T depends on z through a log logistic regression model whose survival function has the form

$$S_T(t|z) = \frac{1}{1 + \lambda t^\alpha e^{\beta z}}, \tag{3}$$

where λ, α and β are some parameters.
Is model (3) an accelerated failure time model? Answer the question in the following ways: i) check if the s.f. of T has the form (2); ii) check if the model for $Y = \log T$ has the form (1) and try to specify the distribution of W. Try to express the parameters λ, α, β in (3) as functions of the parameters β_0, β_1, σ.
(c) Is the failure time model specified by (3) a proportion hazards model? Why?

2. The data in "p42.txt" contain survival times for leukemia patients. The times are in weeks from diagnosis and there are two covariates: white blood cell count (wbc) at diagnosis and a binary covariate AG that indicates a positive or negative (positve=1, negative=0) test related to white blood cell characteristics.
(a) Assess graphically how the survival time is affected by the individual covariates, AG, wbc or even functions of wbc. Try Kaplan-Meier plot or scatter plot, or other graphical exploration if appropriate.
(b) Fit a parametric regression model to the data with appropriate covariates.

(c) Carry out residual analysis to assess the fit of your model.
(d) Based on the final model, interpret the estimated regression coefficients and explain how the survival time depends on the covariates.

3. The bladder cancer recurrence data in "bladder.txt" are collected on a tumor recurrence study for patients with bladder cancer. Individuals had 0 to 4 recurrences during the follow-up time. We will focus on the time to first recurrence, measured from entry to the study. Define the response (failure/censoring) time based on follow-up time ("futime") and first recurrence time (in months) as follows. Define the "failure" time of a patient as his time to first recurrence, if observed. If a patient experienced no recurrences, he is considered "censored" at the end of the follow-up time. The covariates include

treatment group: 1=placebo, 2=drug thiotepa
size: size of the largest initial tumor (in centimeters)
number: number of tumors at initial diagnosis.

Study how these covariates affect the time to first recurrence. Apply the Cox regression model or models to analyze the data. Support your analysis with appropriate graphical exploration (**KM plot** for comparing the treatment groups), model checking and residual analysis (using **deviance residuals** only). Describe and summarize your data analysis and clearly state your conclusions.

4. **This problem is for Stat 886 students only.** Comparison of survival distributions for multiple groups. The data set "BoneMarrowDFTime.txt" given below is extracted from a study of bone marrow transplant for leukemia patients reported by Copelan et al. (1991). The 137 leukemia patients were recruited from four hospitals in the United States and Australia. The patients were treated with a preparative regimen and then given the transplantation. They were grouped into 3 risk categories based on their disease status at the time of transplantation. They were then followed up for up to 7 years. The data set has 3 columns,
1) disease-free survival time (DFtime), which is the time (in days) to relapse, death or end of study,
2) indicator of disease-free status (DFstatus), which =1 if a patient died or relapsed, and =0 if the patient was alive and disease-free, and
3) disease group, which categorizes patients into 3 groups including
acute lymphoblastic leukemia (ALL), identified by Group=1,
acute myeloctic leukemia (AML) and low-risk first remission (Group=2), and
AML high-risk (Group=3).
(a) Based on the method you described in Problem 5 of Assignment 3, carry out a log-rank test to see if the disease-free survival time distributions are the same for the 3 disease groups. Clearly give null and alternative hypotheses, test statistic and its asymptotic distribution.
(b) Also carry out a test to see if the disease-free survival distributions are the same for the 3 groups using a Cox regression model (using LR or Wald test, whichever is appropriate). Check if the Cox model is appropriate for the data, by examining the proportional hazard assumption through residual analysis, and by a graphical comparison of the estimated survival functions obtained from the KM method and the Cox model.
(c) Suppose $\hat{S}_0(t)$ is the estimated survival function for the baseline group (Group=1, or ALL group). Based on the Cox regression model, give expressions for the estimated survival functions for the other two groups.