STAT 486/886 (Winter 2022)

Assignment 2

The assignment is due on Feb. 17 (Thursday), 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. *Nonparametric Log-Log Based Confidence Interval for $S(t)$.* **This problem is for Stat 886 students only.** In a study, the failure times of n subjects (from the same population) are observed subject to right censoring. Let $a_1 < a_2 < \ldots < a_K$ denote the distinct failure times recorded in the study. We have derived $\hat{S}(t)$, the Kaplan-Meier estimate for the survival function $S(t)$ of the underlying failure time distribution, and the Greenwood's formula for estimating the variance of $\hat{S}(t)$. Show that

$$\hat{\text{Var}}\{\log[-\log \hat{S}(t)]\} = \frac{\sum_{j:a_j<t}\{d_j/[n_j(n_j-d_j)]\}}{[\sum_{j:a_j<t}\log(1-d_j/n_j)]^2}$$

where d_j is the number of failures at time a_j and n_j is the number of subjects at risk at a_j. Also describe how you will obtain the log-log based (point-wise) confidence interval for $S(t)$.

2. *Parameterization for Weibull and Exponential Failure Time Models in R.*
Let T be the failure time of a subject. It is known that

$$\log T = \mu + \sigma W, \tag{1}$$

where the parameters $-\infty < \mu < \infty$, $\sigma > 0$, and W has a standard extreme value distribution whose survival function is $S_W(w) = e^{-e^w}$ for $-\infty < w < \infty$.
(a) Show that the failure time T has a Weibull distribution with survival function $S_T(t) = e^{-\lambda t^\beta}$ with parameters $\lambda > 0$, $\beta > 0$. How are μ and σ related to λ and β?
(b) With $\sigma = 1$ in (1), show that T then has an exponential distribution.
Note: These results imply that exponential distribution (for T) is a special type of log

location-scale distribution, with location parameter μ and scale parameter $\sigma = 1$.

3. *Likelihood Method for Weibull Model in R.* Let T be a failure time following a Weibull distribution. Consider the log failure time $Y = \log T$. It is known that Y has an extreme value distribution with survival function

$$S_Y(y) = e^{-e^{\frac{y-\mu}{\sigma}}},$$

where $-\infty < \mu < \infty$ is the location parameter and $\sigma > 0$ is the scale parameter. In the "survreg" function in R, however, the Weibull distribution for T (or extreme value distribution for Y) is expressed with parameters μ and $\phi = \log \sigma$. Assume that failure times of subjects under study arise from Weibull distribution. Let $x_1, \ldots, x_n$ be the observed failure or right censoring times for n subjects. Each subject i ($i = 1, \ldots, n$) has a censoring indicator δ_i, taking values 1 if x_i is a failure time and 0 if x_i is a right censoring time. Consider the log time scale $y_i = \log x_i$. The observed data consist of n pairs $(y_1, \delta_1), \ldots, (y_n, \delta_n)$.
a) With parameters μ and ϕ, write out the likelihood function $L(\mu, \phi)$ for the observed data.
b) Derive the score functions for μ and ϕ.
c) Show that the observed information matrix

$$I(\hat{\mu}, \hat{\phi}) = \begin{pmatrix} \sum_{i=1}^n e^{-2\hat{\phi}} e^{\hat{w}_i} & \sum_{i=1}^n e^{-\hat{\phi}} \hat{w}_i e^{\hat{w}_i} \\ \sum_{i=1}^n e^{-\hat{\phi}} \hat{w}_i e^{\hat{w}_i} & d + \sum_{i=1}^n \hat{w}_i^2 e^{\hat{w}_i} \end{pmatrix}$$

where $\hat{\mu}$, $\hat{\phi}$ are the maximum likelihood estimates of μ and ϕ, $\hat{w}_i = (y_i - \hat{\mu})e^{-\hat{\phi}}$ and $d = \sum_{i=1}^n \delta_i$. Explain how you will estimate the variance matrix of $(\hat{\mu}, \hat{\phi})^T$.

4. The following data consists of the times (in months) from relapse to death of 9 leukemia patients after bone marrow transplantation. Starred observations are censoring times.

6, 4, 6, 3, 9*, 9, 10*, 13, 11*

(a) Assume that the survival time (from relapse to death) has an exponential distribution with hazard rate λ. Construct the likelihood function for λ. Find the maximum likelihood estimate (m.l.e.) of λ and the m.l.e. of the mean survival time for the patient population.
(b) Calculate **by hand** the K-M estimate of the survival function $S(t)$ and the variance estimate of $S(t)$.
(c) Find the nonparametric estimate for the mean survival time of the patient population. Does this agree with the corresponding parametric estimate you find in (a)?
(d) Use R or SAS to plot the KM estimate of the survival function. On the same graph, plot also the estimated survival function based on the exponential model in (a). Does the exponential model in (a) fits the data well? Why?

5. One of the goals of recent research on antiretroviral therapy for treating HIV-infected patients is to compare the new triple-drug combinations to the existing two-drug regimens. The triple-drug regimens are expected to maximize antiviral activity, maintain long-term efficacy and reduce drug resistance. A randomized study was conducted to compare a two-drug regimen (AZT + zalcitabine (ddC)) versus a triple-drug regimen (AZT + ddC + saquinavir).

The data below give the time T from administration of treatment (in days) until the CD4 count (a type of white blood cell measure indicating progression of HIV disease) reached a prespecified level for the two groups. The asterisks (*) denote censoring times.

Two-drug group: 85 32 38* 45 4* 84 49 180* 87 75 102 39 12 11 80 35 6
Triple-drug group: 22 2 48 85 160 238 56* 94* 51* 12 171 80 180 4 90 180* 3

The data is also given in the file "hiv.txt". Use R or SAS to analyze the data and answer the following questions.
(a) Obtain the Kaplan-Meier estimate and the Nelson-Aalen estimate of the survival functions for two groups respectively, and plot them. Are the two treatment regimens different in affecting the progression of HIV disease?
(b) Based on the K-M method, for each group (two-drug, or triple-drug), estimate the median of time T. For each group, also find the approximate 95% confidence intervals for the survival probability at 60 days. Find both the plain confidence intervals, and the log(-log)-based confidence intervals.
(c) For the two-drug group, plot the N-A estimate of the cumulative hazard function for T. Comment on the shape of the estimated $H(t)$. Would it be reasonable to model T by a Weibull distribution? (No need to fit the model.)