Survival Analysis
Biostatistics 140.641

Computing Homework – Due Date: Tuesday, 10/25/2022

Data Set: PBC data

Description of data. A total of 424 patients with Primary Biliary Cholangitis (PBC), referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The outcome variable, T, is time from registration to death. The first 312 cases in the data set participated in the randomized trial, and contain largely complete data. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so there are data here on an additional 106 cases as well as the 312 randomized participants. Missing data items are denoted by ”.”. At least one space separates each variable in the data file. Censoring was due to liver transplantation for twenty-five subjects with the following case numbers: 5, 105, 111, 120, 125, 158, 183, 241, 246, 247, 254, 263, 264, 265, 274, 288, 291, 295, 297, 345, 361, 362, 375, 380, 383. (See file “pbc data code.doc” for data codes)

Questions

1. (50%) Suppose the failure time variable T has the Exponential(θ) distribution, where $\theta > 0$ is the constant hazard parameter in the Exponential distribution. Let $Y = min(T, C)$ be the observed time and $\Delta = I(T \leq C)$ the censoring indicator. The observed data include $(y_{11}, \delta_{11}), ..., (x_{1n}, \delta_{1n})$, which are realizations of i.i.d. $(X_{11}\Delta_{11}), ..., (X_{1n}, \Delta_{1n})$ from the placebo/control arm, and $(z_{21}, \delta_{21}), ..., (x_{2m}, \delta_{2m})$, which are realizations of i.i.d. $(X_{21}, \Delta_{21}), ..., (X_{2m}, \Delta_{2m})$ from the DPCA/treatment arm.

   (a) Write out the likelihood function for θ based on observed data from each arm of the clinical trial. Do you need any assumptions when you write out your likelihood functions?

(b) Derive the MLE (in its theoretical form) for θ for each arm and a variance estimate (also in its theoretical form) of the MLE in terms of $\{(x_{1i}, \delta_{1i})\}_{i=1}^{n}$ and $\{(x_{2j}, \delta_{2j})\}_{j=1}^{m}$.

(c) On the basis of the PBC dataset, use your MLE to estimate the hazard parameters θs of both arms and their associated confidence intervals.

(d) On the basis of Exponential(θ) distribution, using the PBC dataset to plot out the survival function estimates for each arm and the associated pointwise confidence intervals based on your MLE.

(e) For the same PBC dataset, plot the Kaplan-Meier curves of both arms and their associated pointwise confidence intervals.

(f) Discuss the comparison between the Kaplan-Meier curves (which are nonparametric estimates) and the survival curves from the exponential distribution model (which are parametric estimates) for each of the two arms.

(g) Can you estimate the medians from the Kaplan-Meier curves? What other summary measures would you like to suggest for the Kaplan-Meier Curves?

2. (50%) Comparing the two arms of the PBC data by the log-rank test. Use ALL the PBC data from 424 patients to conduct the Cox proportional hazards analysis. Attach your computer outputs as appendix if you think they are helpful.

(a) (Hypothesis testing for the two arms) Performing log-rank test to test the difference between the D-penicillamine (DPCA) and the placebo groups. State the null hypothesis. Is the use of the log-rank test appropriate for analysis? Discuss your results.

(b) (Assessment of treatment efficacy) Use ALL the PBC data from 424 patients and apply the Cox proportional hazards model to estimate the hazards ratio between the D-penicillamine (DPCA) and the placebo. State your model and provide your model estimates. Based on your estimates, justify if DPCA significantly improves patient's survival.

(c) (Study of natural history) Use ALL the PBC data and apply the Cox proportional hazards model to study the association of patient's survival with the following demographical and prognostic factors: Age, Albumin, Alkaline Phosphatase, Ascites, Bilirubin, Edema, Hepatomegaly, Platelets, Prothrombin Time, Sex and Spiders. Report your regression coefficients, standard errors and p-values. What would you conclude based on your estimates?

(d) In the data analysis you treat liver transplantation as the source of censoring. Do you think this is appropriate? Does it violate the independent censoring assumption (which is required in regression analysis)?