Survival Analysis

HOMEWORK I

General Guidelines: although you may work in groups on this homework assignment, you must write up your own final assignment. Copying another student's homework or output is not allowed, and each student must run their own programs. In general, computer output and code will not be reviewed unless it is specifically requested as part of the assignment, and the appropriate portions of any output should be inserted directly into your homework solution (rather than attached output at the back). However students may find it useful to keep a copy of their code/output for their own documentation.

1. Exponential Density and Survival-related Functions

The following data consists of the times to relapse and times to death of 10 bone marrow transplant patients, who were followed for up to 45 months after their transplant. Patients # 7-10 were alive and free of relapse at the end of the study. Patients # 4-6 relapsed, but were still alive at the end of the study.

Patient	Relapse Time (months)	Death time (months)
1	5	8
2	8	12
3	12	15
4	20	33+
5	32	45+
6	27	28+
7	16+	16+
8	17+	17+
9	19+	19+
10	30+	30+

(a) Using the data above, calculate the maximum likelihood estimator of the parameter λ for time to relapse and time to death assuming an exponential distribution, $f(t) = \lambda\, e^{-\lambda t}$. Write a brief sentence interpreting this parameter.

(b) Now you will see how powerful this single parameter can be! Using this parameter estimate (round to 3 decimal places), estimate the following quantities:

(i) The mean time to relapse and mean survival time after bone marrow transplant.

(ii) The median time to relapse and median survival time after bone marrow transplant.

(iii) The one-year and two-year probabilities of remaining relapse-free and surviving: in other words, $S_R(12)$ and $S_R(24)$ for relapse and $S_D(12)$ and $S_D(24)$ for death.

(iv) The cumulative probabilities of relapse and death by one and two years (based on the CDF, $F(t)$)

(iv) Based on the exponential distribution with $\hat{\lambda}$ as calculated in (a), calculate the conditional probability of being relapse-free after 2 years given that one has remained relapse-free for at least one year. How does this compare with the probability of remaining relapse-free one year after bone marrow transplant calculated in part (iii)?

(c) If we decide that an exponential distribution is not appropriate and want to estimate the survival distribution non-parametrically, is it possible to estimate the median time to relapse? Is it possible to estimate the median time to death? If so, provide the appropriate estimates.

2. **Kaplan-Meier Survival Estimate**

Triple drug regimens have been used to treat HIV-infected patients because they were expected to maximize antiretroviral activity while reducing the incidence of drug resistance. In one study, investigators compared antiretroviral-naive (eg., newly infected) patients randomized to either a 2 drug regimen (zidovudine (ZDV) + zalcitabine (ddC)) as compared to a 3 drug regimen of ZDV+ddC + saquinavir (SQV). The time to event was measured in days from randomization, and the outcome was response to treatment as reflected by a CD4 count>300 cells/mm^3 (all patients started the study with lower CD4 counts, reflecting poor immune status). The times are shown below only for the 3-drug arm, with "+" indicating censored subjects:

$22, 2, 48, \mathbf{80}, 160, 238, 56+, 94+, 51+, 12, \mathbf{161}, 80, 180, 4, 90, 180+, 3$

(a) Using the above data (and assuming non-informative censoring as necessary), calculate the Kaplan-Meier estimate of the survival function, $\hat{S}(t)$, by hand. Summarize your calculations in a table with columns for events (d_j), censoring (c_j), and risk sets (r_j) at each time t_j.

(b) Repeat the above estimation of $\hat{S}(t)$ using any software you choose. Also calculate pointwise 95% confidence intervals for $\hat{S}(t)$ using the "log-log" approach and the linear approach. Do either of the approaches result in lower or upper confidence bounds outside the [0,1] interval?

(c) Plot the estimated survival function $\hat{S}(t)$ and pointwise 95% confidence intervals, by hand or using any statistical software package.

(d) Provide the estimated median survival, along with the estimated 25th and 75th percentiles (when possible). Indicate where these percentiles fall on your KM plot from (c) by drawing horizontal lines. What are the actual KM survival estimates corresponding to each of these estimated percentiles?

(e) Calculate the estimated cumulative hazard rate, $\hat{\Lambda}(t)$ at each time t using the Kaplan-Meier survival estimate from part (a) or (b) as the basis.

(f) Calculate the estimated cumulative hazard, $\hat{\Lambda}(t)$, using the Nelson-Aalen estimator.

(g) Plot (i) the estimated cumulative hazard $\hat{\Lambda}(t)$ vs t and (ii) the estimated log cumulative hazard $\log \hat{\Lambda}(t)$ vs $\log(t)$ and use these plots to comment on the appropriateness of the Exponential and Weibull models for this data.

(h) Using the results from part (f), calculate the alternative Fleming-Harrington estimator of the survival function, $\hat{S}_{FH}(t)$ and comment on its agreement with the Kaplan-Meier estimate from parts (a)/(b).

3. Lifetable (Actuarial) Survival Estimate

Now group the data from Question #2 into approximate 1-month intervals (eg., 30-day intervals). Note that the data above is given in terms of days, so 1-month intervals are 0-30, 30-60, 60-90, etc.

(a) Using the grouped data, calculate the actuarial estimate of the survival function for response to antiretroviral treatment (by hand or by computer).

(b) Calculate the estimated hazard function at the midpoint of each time interval and plot.

(c) What can you say about the hazard for treatment response over time? Does an exponential model seem appropriate for this data?

4. Non-Informative and Informative Censoring

Read the articles posted on the course website by Roberts et al. (NEJM, 2004). We define T_1 as the time from placement on waiting list to 'transplantation'. What are the the definition of starting point and end point of the survival time T_1. Would you consider the censoring mechanism non-informative for T_1? Would you consider the censoring mechanism non-informative for the time until a transplant with at least one HLA mismatch becomes available?