BIOST/EPI 537
SURVIVAL DATA ANALYSIS IN EPIDEMIOLOGY

Winter 2025
Instructor: Ting Ye

PROBLEM SET 1

Due by 11:59pm on Friday, January 24th, 2025

PROBLEM 1 (adapted from Klein and Moeschberger, 2003)

A large number of healthy individuals were enrolled in a study beginning January 1, 1970, and were followed for 31 years to assess the age at which they developed breast cancer. Individuals had clinical exams every 3 years after enrollment.

(a) Discuss the type of truncation present in the study.

Furthermore, for four selected individuals described below, discuss the types of censoring present.

(b) Individual A enrolled in the study at age 32 and never developed breast cancer during the study.

(c) Individual B enrolled in the study at age 40 and was diagnosed with breast cancer at the fifth exam after enrollment.

(d) Individual C enrolled in the study at age 56 and died from heart failure at age 61, without ever being diagnosed with breast cancer. Post-mortem examinations confirmed that this patient never developed breast cancer.

(e) Individual D enrolled in the study at age 43 and moved away from the community at age 57, without ever being diagnosed with breast cancer by investigators.

If instead of age at onset of breast cancer we were interested in studying the time from enrollment into the study until onset of breast cancer, would your answers above be any different? If so, how?

PROBLEM 2.

You wish to study the distribution of time from diagnosis of Crohn's disease, a chronic inflammatory condition of the intestinal tract, until first bowel resection surgery or death. Suppose that you have access to a national registry of all current inflammatory bowel disease patients. On a particular date (i.e., the day of study sampling), you enroll a random collection of Crohn's disease patients currently in the registry and who have not yet had bowel resection surgery. The date of diagnosis is available for all patients, and you prospectively follow all recruited patients for 10 years.

Suppose that A represents age at diagnosis, and that T is the time from diagnosis until either bowel resection surgery or death, whichever occurs first.

(a) Describe the types of censoring and truncation affecting the data from this study.

(b) Consider the following statement: in this study, while T is subject to truncation, A is not and hence the observed ages at diagnosis are not affected by selection bias. Do you agree with this statement? Briefly explain why.

PROBLEM 3.

Read Dickson et al. (1989) to learn about the Mayo PBC dataset, and answer the following questions about that study:

(a) What is the study population?

(b) What is the initiating event (or time zero)?

(c) What is the terminating event?

(d) What is the time scale?

(e) What are the causes of censoring?

(f) For each of the causes above, comment on whether you believe the underlying censoring mechanism may be related to the outcome of interest?

(g) Figure 3 in Dickson et al. (1989) presents survival curves for 106 cross-validation patients. Based on the Kaplan-Meier curves you see on this plot, for each of the three risk groups, approximately what proportion of individuals die **within five years**?

PROBLEM 4.

Investigators followed a cohort of 238 individuals with heroin use disorder who entered methadone maintenance programs in either of two clinics in Sydney, Australia over an 18-month period during the late 1980s. The purpose of the study was to identify factors associated to retention in methadone maintenance. Time from entry into the study to exit from methadone maintenance is the duration of interest. Details can be found in Caplehorn and Bell (1991).

In this problem, you will perform a parametric analysis of data from this study. The dataset in question, called `methadone.csv`, can be found on the canvas website and loaded using the `read.csv()` function. The variables in this dataset are:

- `id`: patient identification number;
- `clinic`: clinic identifier;
- `event`: indicator of exit from maintenance;
- `time`: observed follow-up time (in days);
- `dose`: maintenance methadone dose (in mg/day);
- `prison`: indicator of previous incarceration.

(a) Compute the average follow-up time and the proportion of censored observations.

(b) Using the available data, fit exponential, Weibull and generalized gamma models to the distribution of time to exit from maintenance. For each model, report parameter estimates, associated 95% confidence intervals and maximum loglikelihood value.

(NOTE: For the exponential model, report on parameter λ. For the Weibull model, report on parameters λ and p. For the generalized gamma model, report on parameters μ, σ and Q.)

(c) Plot the survival function corresponding to each of the above parametric fits as well as a nonparametric estimator of the survival function on the same graph, and comment visually on the adequacy of the models considered.

(d) Is the Weibull model an appropriate simplification of the generalized gamma model in this example? Justify your answer by performing an appropriate statistical test.

(e) Using a Weibull model, provide an estimate and 95% confidence interval of:

i. the median time until exit from maintenance;
ii. the probability that no exit will occur by one year;
iii. the probability that no exit will occur by two years given that no exit has occurred by one year.

(f) Is the exponential model an appropriate simplification of the Weibull model in this example? Justify your answer by performing an appropriate statistical test.

(g) Separately fit an exponential model to the subset of individuals in clinic 1 and clinic 2. Report parameter estimates and corresponding 95% confidence intervals. Use the output of these two fits to determine whether the distribution of time to exit from maintenance differs significantly by clinic. Justify your answer by performing an appropriate statistical test.

(h) Repeat the last problem but substituting clinic by history of incarceration (i.e., prison).