Assignment #1 (Due Tuesday, Feb 19)

SOCY 7717 Event History Analysis and Sequence Analysis

Wen Fan

February 4, 2019

- *Only hard copies will be accepted.*
- *This assignment, like all others in this course, should be answered in complete English sentences. You also need to turn in the log file (not the do file) as part of your homework assignment. Please keep the log file clean. In particular, do not include incorrect codes, error messages, or codes and outputs that are irrelvant to the questions, unless they are related to some point you are raising in your discussion.*
- *Limit the length of your report for the assignment to 3 pages, including text, tables, figures, and any references. Please embed tables and figures within the text near to where they are discussed rather than at the end of your report. The font size should be at least 11.5 pt, and the margins should be at least 1 inch on all sides.*
- *If you need to present model estimates, please present them in a formatted table as you would for a journal article (not outputs copied directly from Stata).*
- *In response to each question, please make it clear which statistical evidence you draw on (e.g., F-test values, p-value) to reach the conclusion; simply presenting outputs lacking intelligible discussion are unacceptable. Remember to take uncertainty into account in your interpretations of coefficients.*
- *These problems need to be worked out independently. All students should be prepared to answer the questions orally in class.*
- *Except for clarification, I will not answer questions directly related to this assignment before it is graded.*

1. A large number of disease-free individuals were enrolled in a study beginning January 1, 1988, and were followed for 30 years to assess the age at which they developed lung cancer. Individuals had clinical exams every 3 years after enrollment. For four selected individuals described below, discuss the types of censoring and truncation that are represented.

(1) A healthy individual, enrolled in the study at age 30, never developed lung cancer during the study. (0.5 points)

(2). A healthy individual, enrolled in the study at age 40, was diagnosed with lung cancer at the fifth exam after enrollment (i.e., the disease started sometime between 12 and 15 years after enrollment). (0.5 points)

(3). A healthy individual, enrolled in the study at age 50, died from a cause unrelated to the disease (i.e., not diagnosed with lung cancer at any time during the study) at age 61. (0.5 points)

(4). An individual, enrolled in the study at age 42, moved away from the community at age 55 and was never diagnosed with lung cancer during the period of observation. (0.5 points)

(5). Confining your attention to the four individuals described above, for each of them write down the output you would get in `Stata` after you `stset` the data (i.e., what are the values of `_t0`, `_t`, and `_d` for each of the four individuals). (0.5 points)

2. The National Longitudinal Survey of Youth (NLSY) is a stratified random sample that began in 1979. Youths, aged 14 to 21 in 1979, have been interviewed yearly through 1988. Beginning in 1983, women in the survey were asked about any pregnancies that have occurred since they were last interviewed (pregnancies before 1983 were also documented).

The data below consist of the information from 927 first-born children to mothers who chose to breast feed their child and who have complete information for all the variables of interest. The outcome variable for this exercise is the duration of breast feeding in weeks. Construct a life table by hand based on the data below (you can use an excel sheet if needed). Note that to get full credit, it should be clear how each number in your results is obtained. (2.5 points)

Weeks	Number lost to follow-up or withdrawn without being weaned	Number weaned
0-1	2	77
2-3	3	71
4-5	6	119
6-7	9	75
8-10	7	109
11-16	5	148
17-25	3	107
26-36	0	74
37-52	0	85
53+	0	27

3. Listed below are lengths of time (in months) staying in foster home for a sample of 14 children. Right-censored times are denoted by a "+" as a superscript.

$1, 3^+, 3^+, 5, 6, 7^+, 7^+, 8, 8, 9, 10^+, 12, 12^+, 12^+$

Using these data, answer the following questions:

(1). Compute the Kaplan-Meier estimate of the survival function by hand. (2 points)

(2). Draw a plot of the KM estimate computed in problem 3(1). (0.5 points)

(3). Find the median survival time. (0.5 points)

4. For this question, download the data on the survival of 314 European cabinet governments from here and perform the following exercise in `Stata`. The variables in the data should be self-explanatory, but key to this exercise are the `durat` variable that records the duration of the government and the `censor` variable that is coded 1 if the observation is censored.

(1). Set up the data for survival analysis. (0.5 points)

(2). Plot the the Kaplan-Meier function along with the Greenwood standard errors, as well as Nelson-Aalen estimator of the cumulative hazard. What is the median survival time? (0.5 points)

(3). Plot the smoothed hazard function, and see how the plot changes as you vary the `width` option. (0.5 points)

(4). Do survivor function rates differ depending on whether or not the government has a numerical majority (the variable `numst`, `1` denotes majority and `0` otherwise). (0.5 points)

5. (Bonus question) Read the paper *Elvis to Eminem: Quantifying the price of fame through early mortality of European and North American rock and pop stars* by Bellis et al. (2012). A copy can be found here.

(1). Explain briefly how Figure 1 is obtained and what information this figure tells you. (0.5 points)

(2). Explain briefly how Figure 2 is obtained. Comment on this figure. (0.5 points)