Assignment #2 (Due Tuesday, March 26)

SOCY 7717 Event History Analysis and Sequence Analysis

Wen Fan

March 6, 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 4 pages, including text, tables, figures, and any references. Your score for this assignment will be determined by the following function: I(length $\leq$4 pages) $\times$ actual score. 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.*

For questions 1-3, use the data from assignment 1 on the survival of 314 European cabinet governments (download here). If not specified otherwise, all models will use the following covariates:

`invest`: whether or not an initial confirmatory vote is required by the legislature

`polar`: percentage of support for extremist parties

`numst`: whether or not the government has a numerical majority

`format`: a count of the number of attempts to form a government prior to the official formation of the government

`postelec`: a binary indicator denoting if the government was formed immediately after an election

1. (1). Estimate the generalized Gamma model for these data. Among the distributions encompassed within the generalized Gamma model, which, if any, provides the best fit to the data? Can we rule out any? Please explain your answer. (0.5 points)

(2). Using the BIC criteria, which model fits best among the Gompertz, log-logistic, log-normal, Weibull, and exponential distributions? (0.5 points)

(3). Can you use likelihood ratio test to determine whether a Weibull model or an exponential model fits the data better? Between a Gompertz model and a log-normal model? For both pairs, if your answer is yes, perform the likelihood ratio test and determine which model is better; if not, explain why. (1 point)

(4). Plot the estimated hazard rate by `numst`, based on your preferred model in question 1(2). Describe the main features of the hazard function as displayed in this graph. (0.5 points)

2. (1). Estimate a Weibull model (in proportional hazards form). What conclusion can you draw based on the ancillary parameter p? (0.5 points)

(2). Interpret the coefficients from the Weibull PH model for `polar` and `numst`. (1 point)

(3). Estimate a Weibull model (in accelerated failure time form). Take the coefficient for `invest` from this AFT model, show how to obtain the coefficient for `invest` in an otherwise similar Weibull PH model. (0.5 points)

(4). Interpret the coefficients from the Weibull AFT model for `format` and `postelec`. (1 point)

(5). Based on the Weibull model, use the Cox-Snell residuals to conduct a diagnostic test. (0.5 points)

3. (1). Estimate a Cox model using the Efron method for ties. Do your inferences change based on results from the Cox model compared with the Weibull model you fit for question 2? (0.5 points)

(2). Evaluate the proportional hazards assumption for `invest` and `postelec` using at least two different methods. Are you confident that the PH property holds for these two predictors? (1 point)

(3). Suppose that the variable `polar` violates the proportional hazards assumption. Describe two ways to address this problem. You don't need to run any analysis in Stata but include relevant syntax either in your log file (as comments, for example) or in your answer to this question. (0.5 points)

(4). Estimate a stratified model by numerical status (again, using the Efron method for ties). Plot a baseline cumulative hazard function for numerical majority and minority groups. (0.5 points)

4. For this question, use the partnership duration data (download here). We've used this data set for a few times in class, but the event being examined here is partnership dissolution. The `time` variable denotes event time in months.

(1). Use the following code to first generate a variable `chdbirtime`, which measures the

time when the respondent's child was born (in months, with the same origin as the event time). Fit a Cox model with a time-dependent variable presence of children, `sex`, and `married`. Interpret the coefficient for presence of children. (hint: you may find this article helpful: Stata tip 8: Splitting time-span records with categorical time-varying covariates) (1 point)

```
set seed 20190306
gen chdbirtime = ceil(time * runiform()) * status + ceil(50 * runiform()) * (1
- status)
```

(2). Are there any influential cases (based on DFBETA)? Briefly describe how you plan to handle influential cases if there are any. (0.5 points)

5. (Bonus question) Suppose that we have a sample of five respondents:

- Observation 1: event time interval-censored (2, 5]
- Observation 2: observed event time at 7
- Observation 3: observed event time at 4
- Observation 4: event time left-censored at 8
- Observation 5: event time right-censored at 3.5

(1). Write down the likelihood function for an analysis of event time. (0.5 points)

(2). Now, consider only observations 2, 3, and 5. Suppose event time follows an exponential distribution. The survival function for an exponential distribution is $S(t) = \exp(-\lambda t)$ and the density function is $f(t) = \lambda \exp(-\lambda t)$, where t indicates time and λ is the rate parameter. Find the maximum likelihood estimator (MLE) for λ. (0.5 points)