

MLE 2: Event History Analysis

Problem Set 1

Overview In this problem set, I want you to get some experience with the Kaplan-Meier estimator, parametric and Cox models.

Directions Please access the data and answer the questions/estimate the models that are given below. The data are available at:

<http://www.u.arizona.edu/~bsjones/eventhistory.html>

If you use **Stata**, the data set has already been `stset`. To help, here is some information on the data set:

Warchest Data

For the following questions, the definition of the variables are as follows (note that these data are used in an article by Jan Box-Steffensmeier in the May 1996 *American Journal of Political Science*. Here are the data:

`_t` denotes the length of time until a “high quality” challenger enters a House race against an incumbent. The scale of the dependent variable is in terms of the number of weeks that pass until a high-quality challenger emerges. The minimum is 1, denoting 1 week; the maximum is 90, denoting 90 weeks. I denote this variable as *Time-to-Entry*. Note that in **Stata**, since the data are already `stset`, you do not need to specify this as the dependent variable as **Stata** will already recognize it as such.

`iv` denotes the prior vote the incumbent received (scaled between 0 and 1; the minimum value is .5 (denoting 50 percent) and the maximum value is 1 (denoting 100 percent)).

`ec` denotes the incumbent’s “warchest”; that is, the amount of money the incumbent has in reserve to use at his or her discretion (scaled in millions of dollars). The minimum value is .00069, which corresponds to \$690; the maximum value is 1.688, which corresponds to \$1,688,000.

`south` is a dummy variable denoting whether or not the incumbent is in a Southern state (1 denotes South, 0 denotes non-South).

`dem` is a dummy variable denoting whether or not the incumbent is a Democrat (0 = Republican).

`_d` is a “failure” indicator. 1 means the event occurred, 0 means the event has not occurred (i.e. it is right-censored).

Militarized Interventions Data

The following are the variable definitions.

`pbal` is the relative capabilities index. Scores on this variable closer to 1 indicate a materials capabilities imbalance in favor of the intervenor state and scores closer to 0 indicate an imbalance in favor of the target state.

`idem` and `tdem` are the policy democracy scores for the intervenor state (`idem`) and the target state (`tdem`). The scores range from -10 (least democratic) to 10 (most democratic).

`ctg` is a dummy variable coded 1 if the two states share territorial contiguity (i.e. a border) and 0 if not.

Questions

1. How many cases are right-censored? What does right-censoring mean in the context of this research problem? (5 points)
2. Estimate the generalized gamma model for these data. Use as covariates the ones discussed above (DON'T include d as a covariate!). Among the distributions encompassed within its CDF, which, if any, provides the best fit to the data? Can we rule out any? If you use Stata, it will report σ instead of p . This doesn't make any difference: $1/\sigma = p$. Since the test is against 1, either *sigma* or p is fine to use. Also in Stata, use the unlogged σ for your hypothesis test. Please explain your answer and show any relevant work. (10 points)
3. Using the AIC criteria, which model fits best among the Gompertz, log-logistic, log-normal, Weibull, and exponential distributions? How do your conclusions here compare to your answer in question 1? (10 points)
4. Plot the estimated hazard rate from your preferred model in question 2. Describe the main features of the hazard function as displayed in this graph (10 points).

For the next set of questions, use the militarized interventions data set.

5. Give the Kaplan-Meier estimates for these data as well as provide a plot of the estimated survival function. What are the main features of this plot? (5 points)
6. What is the median survival time for these data. Note that the percentiles of the survival times are computed by the following formula:

$$\hat{t}(\text{p'tile}) = \min\{t_j \mid \hat{S}(t_j) \leq 1 - (\text{p'tile}/100)\},$$

where $\hat{t}(\text{p'tile})$ is the estimated survival time percentile and t_j is the minimum j th ordered failure time conditional on $\hat{S}(t)$ being less than or equal to the specified probability of the

percentile. (That is, the median is taken to be the smallest observed survival time such that $\hat{S}(t) < .5$.) Provide a substantive interpretation of this result. (5 points)

7. Estimate a Weibull (in proportional hazards form). Substantively interpret each covariate as well as the ancillary parameter (interpret the ones that you might deem “insignificant” as well). (10 points)

8. Estimate a Cox PH model. Substantively interpret each covariate. Do your inferences change based on results from the Cox model compared to the Weibull? (10 points)

9. Using a link test, evaluate the proportional hazards property for this model? Following this, estimate covariate specific tests and the global test of proportionality for this model. What are your conclusions from these tests? Are you confident the PH property holds for this model? (10 points)