Workshop on Survival Analysis
Frederick J. Boehmke
Computer Lab #1: Discrete EHA

First we will learn about how to estimate and interpret a discrete EHA model. Doing so involves going from information about the timing of events for each unit to setting up the data with the proper structure to estimate an EHA model. We will analyze U.S. state data on the diffusion of a single policy, here hate crime laws, with some standard control variables.

We will then explore different ways to account for duration dependence in a discrete EHA: linear and polynomial functions of time, time fixed effects, and linear and cubic splines.

Part I

The included data set includes adoption dates (if adopted) for hate crime laws across the American states from the first case in 1978 through the last observed adoption (from the source data) in 2003. It also includes information on the states for the relevant time period to use as independent variables. I've included information from the adoptions to facilitate creating the dependent variable for EHA.

1. Open the included data set, `exercise01discrete.dta`. As always, use `describe`, `summarize`, `tabulate`, and `table` to explore the basic features of the data.
2. Create the dichotomous dependent variable with a zero for years in which a state is at risk but does not adopt, a one in the year of adoption, and missing for years following adoption. These data were observed through 2003, so that will mark the last year of the risk set. For states that do not adopt, the dependent variable will stay zero until the end of the risk set (they may adopt after 2003). Remember that in Stata missing values are treated as infinity, so be careful using inequalities and remember the `missing()` function.
3. Create graphs of the number of adoptions per year through 1999 and then the survival function each year. An easy way to do this relies on versions of the `graph bar` command.
4. Run a basic logit EHA model controlling for citizen ideology, total population, and real income per capita.
5. Then try probit and a cloglog models to compare.

Part II

Now we can introduce various controls for duration dependence and compare.

1. Create a variable for a linear time trend, t, starting with $t = 0$ in the first year of the risk set. Run the logit models and save the results.
2. Now generate t^2, add it into the prior logit model for a quadratic time trend, and save the results.
3. Now generate t^3, add it into the prior logit model for a cubic time trend, and save the results.
4. Compare the results of these four models. (Hint: `estimates store` and `estout` or `estimates table` are your friends here.)
5. Now generate time fixed effects and run the model with those.
   `.xi i.year, noomit`

Part III [Illustration]

This section goes into details on using linear and cubic splines to control for duration dependence. We will likely skip over it, but I will demonstrate if possible and leave you with the code for later exploration.

1. Generate a linear spline at tertiles using `mkspline` and run the model with those.
2. Compare all your results.
3. Generate a cubic spline with 4 knots using `mkspline` and run the model with those.
4. Compare all your results.
5. **[Optional]** Graph the time effect implied by each. I would collapse the data to the mean of all the substantive and time-related variables by year. Then replace the substantive variables with their mean values by using `egen` to get the mean in a new variable and `replace` to fill in it to the variable included in the model. Then you may use `predict` to estimate the outcome. It helps to compare in both the latent variable scale (Y^*) and in the probability scale. You can start with one model. If you used `estimates store` earlier, you can `estimate restore` now to go through each model after you collapse.