Workshop on Survival Analysis
Frederick J. Boehmke
Exercise #3: Non-proportional Hazards

In this exercise we will investigate tests for the proportional hazards assumption and then ways to address and interpret it. This will also provide an opportunity to generate and include time-varying covariates as part of the correction.

We will use data about position taking by members of Congress on NAFTA from Box-Steffensmeier, Arnold, and Zorn (1997). The data indicate the number of days before the NAFTA vote that each member of Congress stated his or her preference on the bill started by President George H.W. Bush and supported by then-President Clinton. The first member announces 463 days before the vote, so that is set as time zero. We will control for a host of factors about each member and his or her district that help determine an early or late announcement.

Part I

First we will run some models, test for NPH, and implement the correction using Stata's built-in syntax.

1. Open the included data set, `exercise03nonph.dta`. As always, explore the basic features of the data.
2. `stset` the data with `timing` as the duration outcome. We will assume all member of Congress declare a position, if only by their vote on the final day.
3. Run the following Cox model:
   ```
   .stcox corptpct labtpct mexbordr dleader rleader ncomact ideol
      pscenter hhcenter, nohr
   ```
4. After running it, run `estat phtest, detail` to test the proportional hazards assumption. If it fails for a variable, use `estat phtest` with the `plot` option to examine the Schoenfeld residuals over time. Repeat for a variable that passes with flying colors to compare.
5. Stata has an option in `stcox` that allows you select variables that do not meet the PH assumption and then to do the correction with redoing the data set. Use `tvc()` in conjunction with `texp()` to allow the effect of the offending variable to change with the natural logarithm of time.
6. Plot the estimated hazard ratio against time. I'd use the `twoway function` command to graph $\exp(\beta_k + \beta_t * \ln(t))$ against time for a variable k that violates the PH assumption. Note that that you will have to specify the equation names for the variable and the time varying component — these are given above the coefficient names in the results.

Part II

Now we will use these data to explore continuous-time duration data with time-varying covariates. We will motivate this by accounting for NPH "by hand", that is by converting the data to TVC structure and including the interaction of `mexbordr` with the natural logarithm of time. This will also facilitate comparing the consequences of implementing this correction incorrectly, i.e., without moving to TVCs.

1. Now let's do this the wrong way by manually adding in an interaction between the offending variable and the log of time (i.e., rather than using the `tvc()` option). How do these results compare? Plot them again.

2. Finally, let's do the fix the right way by hand. To do this we have to expand the data set to have multiple observations for each unit to capture how the TVCs evolve over time. Since this is a Cox model, we only need data for units in the risk set at the observed failure times. We can use `stsplit` to create this easily, but we'll have to re-`stset` the data with an identifier first.[1] You'll also need a failure indicator since we are treating these as multiple records per subject now. Once you've re-`stset` the data, rerun the base Cox model and make sure the results match.

3. Now you can `stsplit` the data. Use the `at()` option to split it just at the failures. Take a moment to explore what this creates. Now rerun the base Cox model (the one we started with) again just to be sure — the number of observations will change but the results should be identical.

4. Create the interaction between the log of time and the offending variable and run the model with non-PH. Compare the results to what we had before. They should match.

5. If you have time, try other functions of time to see if the natural logarithm is indeed best.

[1] See `help tvc note` for documentation on this.