Workshop on Survival Analysis
Frederick J. Boehmke
Exercise #5: Time-Varying Covariates

In this exercise we will learn how to set up and run survival models with time-varying covariates. We will again use our NAFTA data so that we can compare the single-record per subject format to the multiple record per subject format needed to have different values of the covariates over time.

We will add time-varying covariates corresponding to the observed support each day. This will allow us to evaluate whether members of Congress take cues from each other.

In this case we will build up the data from its single record form to have a cross-sectional time-series structure which one observations per member per day. A new variable will mark the day of failure. We will then `stset` the data in this new form and estimate our models.

Part I

Let's open the data, `stset` as before and run the single failure model to set a baseline. Then we'll add some variables to track subjects and re-`stset` the data. Finally, we'll modify the data to have multiple records per subject, and again `stset` and estimate our model.

1. `stset` the data with `timing` as the dependent variable just as before.
2. Run the single failure Weibull model we used in the previously:
   ```
   .streg corptpct labtpct mexbordr dleader rleader ncomact ideol
     pscenter hhcenter, time dist(weibull)
   ```
3. We need to create some variables to track observations and failures for when we have multiple records per subject.
   ```
   .generat congid = _n
   .generat day_ann = timing
   .generat failure = 1
   ```
4. Now `stset` the data with `timing` as the dependent variable but also referencing the id and indicating that all subject fail:
   ```
   .stset timing, id(congid) failure(failure)
   ```
5. Estimate the Weibull model again. It should produce identical results.

Part II

Now let's create time-varying data (though no time-varying covariate just yet).

1. Right now we have 1 observation per subject. We want one per subject per day, so let's create 462 copies of each observation using `expand`.

2. Now we need a variable to mark each day from 1 to 463.

3. Finally, we need a variable that indicates the day of the announcement. Create an indicator equal to 0 for days before the announcement, 1 on the day of, and missing on days after the announcement. A tab should produce 435 ones.

4. Since we only have one failure per subject, replace the failure variable with missing values on days after the announcement. This effectively marks the risk set.

5. Now `stset` the data with `day` as the new dependent variable. We must reference the id, indicate the failure event, and indicate that all subject fail:
   ```
   .stset day, id(congid) failure(announce) exit(failure)
   ```

6. Estimate the Weibull model again. It should produce identical results.

Part III

Now let's create a time-varying covariate to take advantage of our new setup.

1. We will create variables counting the number of supporters and opponents that have declared prior to each day. This will take a few steps.

   (a) Generate separate indicator variables capturing favorable and unfavorable announcements on the day of those announcements. There should be 435 total ones between them.

   (b) Now get the total such announcements per day for each using `egen()`'s `sum`function.

   (c) Now we want to get a running sum. I use the lag function to accumulate this for each member, which requires me to `xtset` the data.

   (d) Now create lagged values and fill in the first day to 0.

2. Now let's create a variable for net support by taking the difference between the two.

3. Estimate the Weibull model with net support.

4. Estimate the Weibull model with total support and total opposition. Compare the results.

5. Generate a graph that plots the hazard at net support equal to -20, 0, and 20.

Part IV [Optional]

If you want to see why it matters, estimate the time-invariant model but include the observed and net support variables measured only on that day. You can do this easily by adding an `if` to the `stset` command so that there is just one observation per member:

```
stset day if day==day_ann
```

Then rerun the three models we ran with the TVC setup and compare. The time-invariant version is wrong because it ignores the fact that the support values took on different values on days prior to the day of announcement which would therefore have changed the hazard on those prior days.