Using Allison’s Recid data set, carry out an analysis of recidivism using a “discrete-time” approach. Measurement of the timing of recidivism is obviously interval censored.

1. Prior to creating stacked risk sets, examine the hazard using the exploratory data analysis tools supplied by Stata. In particular, obtain a graph of the (smoothed) hazard. [Be sure to save all graphs as files.] Compare your graph with Allison’s Output 3.16:59. Are they quite similar? If not, try different choices of bandwidth to see whether or how the estimated shape of the smoothed hazard changes.

2. The employment status variable is time-varying. If you are using Stata, you will find it efficient to use the “reshape strategy” illustrated in the handout titled “Discrete time survival analysis using time varying covariates: an example of the use of -reshape- to create a stacked data set.”

3. No events occur during weeks 29, 41, and 51. For a discrete-time approach, we still need risk sets for those weeks. Fortunately, we do not need to go out of our way to create these risk sets, because application of the reshape command to the employment status variable solves the problem. You should verify this.

4. Allison’s Output 5.7:131 displays the results of a Cox regression of recidivism on a set of covariates (excluding employment status). Fit a corresponding complementary log-log regression allowing risk set to enter the regression as a set of dummy variables. Comment in response to each of the following: (a) How does Stata handle weeks 29, 41, and 51? (b) How closely do the complementary log-log results match the SAS-estimated Cox regression? (c) What happens if you restrict the coefficients of weeks 28-29 to be equal, and do likewise for weeks 40-41 and 50-51? Why, why, and why?

5. Most researchers would find it difficult to make sense of a year’s worth of coefficients for week-specific risk set dummies, relying only on visual inspection. Restrictions on the coefficients are likely to be helpful. Consider several alternatives. At the simplest extreme, one can impose linearity on the hazard. More complex alternatives include nonlinear smoothers and piecewise linear splines. Even if you try a nonlinear smoother, also try piecewise linear splines. Stata makes this easy with the “mkspline” command. You may find it helpful to consider one set of splines with knots at every fourth week, and another set with knots at every 13th week. For each set of splines, graph the fitted baseline hazard. To aid your deliberations, make repeated use of Stata’s “lrtest” command, thinking carefully about which regressions are nested. Out of this work, settle on a shape for the baseline hazard, and defend it.

6. Allison’s Output 5.13:145 extends the list of regressors used in Output 5.7 with the inclusion of the employment status variable. Replicate Allison’s Cox regression results, using complementary log-log regression. To do this, you will need to revert to the weekly risk set dummies.
7. At the bottom of page 145, Allison reports what happens when the employment status variable is included lag-1. Allison uses the lagged employment status variable to solve a potential endogeneity problem. It is useful to know how to create the correct lagged variable in this kind of circumstance. Here’s how, in Stata:

- Let the ID variable be named “id”, the risk set variable be named “sequence”, and the employment status variable be named “emp”.
- The lagged variable can be created using

```stata
sort id sequence
by id: gen emplag = emp[_n-1]
```

8. Using a complementary log-log regression, replicate Allison’s Cox regression results for the lagged employment status variable.