Data-Analysis Exercise

Fitting and Extending the Discrete-Time Survival Analysis Model

(ALDA, Chapters 11 & 12, pp. 357-467)

Purpose of the Exercise

This data-analytic exercise supplements our presentation on specifying, fitting and extending the discrete-time survival analysis. The immediate purpose for the exercise is to "open the door" to your implementation of discrete-time survival analyses with the *SAS* statistical package, by asking you to program and execute a series of simple statistical analyses that replicate, and extend, the core material contained in *ALDA* and featured in our presentation.

General Resources for Conducting the Exercise

While we provide self-contained instructions below for conducting the data-analysis exercise, we recommend that you keep the following resources close at hand:

- *ALDA* itself, and the handouts from our presentation.
- The *UCLA Academic Technology Services* website that supports *ALDA*, at http://www.ats.ucla.edu/stat/examples/alda/ . This site contains downloadable versions of the datasets used in *ALDA*, specimen computer programs (in *HLM*, *MLwiN*, *S+*, *SAS*, *SPSS*, *Stata*) and sample computer output for all exhibits in the book, using each of the software programs.

Datasets

The dataset used in this exercise has been formatted by *UCLA Academic Technology Services* and is available for download, both as text and system files from their website. For our current purposes, the dataset has already been downloaded onto your workstation and stored in a directory on the hard drive. You will need to know the address of this directory in order to conduct the analyses below.

Data-Analysis Exercise

Please proceed sequentially:

1. All of the tasks in this exercise use the data on "grade at first heterosexual intercourse" in *ALDA* Table 11.1 on p. 360. Please make sure that you have this exhibit close at hand.
2. The files that contain these data have the prefix *firstsex*. They appears in the data directory several times with different suffixes denoting different versions of the file -- as raw text, in person-level and person-period format, and as system files appropriate for analyses with the different software packages.

a. First, examine the *raw text person-level* dataset, *firstsex.txt*, by opening *Windows Explorer* and clicking on the filename. Windows *Notepad* should then display it. Make sure that the first two lines of the dataset match the following:

```
"ID","TIME","CENSOR","PT","PAS"
1,9,0,0,1.978866614548087
```

b. Notice that this data-file contains comma-delimited text, with a first line that provides the names of the variables it contains in quotes: *ID*, *TIME*, *CENSOR*, *PT*, *PAS*. By consulting *ALDA*, match these variable names with the variables description on p. 358 ff. Make sure that you can identify the values of each of the variables for the first case in the dataset, and that you know what each represents.

c. Now, in a similar fashion, examine the *raw text person-period* dataset of the same data – it is called *firstsex_pp.txt*. Check that the first four lines of the dataset are:

```
"ID","PERIOD","EVENT","D7","D8","D9","D10","D11","D12","PT","PAS"
1,7,0,1,0,0,0,0,0,0,1.978866614548087
1,8,0,0,1,0,0,0,0,0,1.978866614548087
1,9,1,0,0,1,0,0,0,0,1.978866614548087
```

d. Notice the switch in variable names between the *person-level* and *person-period* datasets – specifically, the appearance in the latter of the time dummies, *D7* thru *D12*, which accommodate the different grades in which the adolescent boy could first experience heterosexual intercourse. Again, make sure that you can line up the names of the variables with their values for the first case across the three periods of his event history and that you understand why they take on the values that they do.

3. Boot up *PC-SAS* on your workstation, by clicking on the appropriate program icon or filename.

   a. You will see the *SAS* opening screen that usually, by default, contains the SAS *Explorer*, *Editor*, *Log* and *Output* windows (the *Output* window may be hidden behind other windows). These windows can be opened by clicking on the labeled tabs in the lower toolbar or by using the *VIEW* menu..
      i. You will use the *Editor* window to compose your program code.
      ii. Your output will appear in the *Output* window.
      iii. Your job log will appear in the *Log* window.

   b. If there is any content currently present in the *Editor*, *Log* or *Output* windows, clear it by going to that window and selecting **Edit>Clear All** on the upper tool bar, or on the menu that pops up when you click the right mouse button.

c. Although *SAS* can accept data entry in all kinds of formats, we will be using data in the form of *SAS system files* in this data-analysis exercise. This has the advantage of streamlining the entry of the data along with the corresponding variable names, etc.

4. Now, begin your discrete-time survival analyses with *SAS*.

   a. Read the *firstsex* person-period system file (*firstsex_pp.sas7bdat*) into *SAS* and conduct the contingency table analyses outlined below to estimate the sample frequencies and proportions in *ALDA* Table 11.1, on p. 360, by typing the following code into the *SAS Editor*:

```
proc freq data="C:/????/firstsex_pp.sas7bdat";
  tables pt*period*event/nopercent nocol;
run;
```

   b. **Where you must replace the *fully-qualified filename* within the quotes on the first line of the program to identify the home of the *firstsex* person-period *SAS* system file on *your own computer***.

      i. In *SAS*, "proc" calls on a particular "procedure" for execution, in this case the "freq" procedure to carry out cross-tabulations and compute frequencies, proportions and the like.

      ii. After the "proc" sentence, there are usually additional sentences. In this example, within the "freq" procedure, we are asking for "tables" of *PT* by *PERIOD* by *EVENT*. Notice how the order in which you list these variables has lead to the particular structure of the output.

      iii. After the slash (/) on the second line of the program, you can include options. In this case, we have chosen options that eliminate unnecessary statistics from the output ("nopercent" "nocol") to reduce clutter. (Try rerunning the program without these options (and the slash), and check out what happens).

      iv. It doesn't matter whether you use upper or lower case text in your program. It doesn't matter where you insert spaces, nor how many. But, **you must end each "sentence" with a semicolon** – omitting the semicolon is the most common error in *SAS* programming.

   c. To execute your program, click on the small "running man" icon on the upper tool bar in *SAS* or select **Run>Submit**.

      i. In the unlikely event that you have not made any programming errors (!), then output from your analyses should appear in the *SAS Output* window. You can move among the windows by using the tabs in the lower toolbar.

      ii. If you have made errors, you will find diagnostic messages in the *Log* window that *may* help you to fix the error (or maybe not!).

1. Edit your program, in the *Editor* window, and re-run.
2. If that doesn't work, "clear all" and start again.
3. If that doesn't work, scream loudly and put your foot through the screen.

d. Once you have successfully obtained output (?), print it out for your records and for the computations below.

5. Examine your output, and identify the "no parenting transition" sub-table (*PT*=0)
   a. Within this sub=table, identify:
      i. The size of the *risk set* in each period (last column on the right, labeled "total").
      ii. The number of events in the period (the penultimate column, labeled "*EVENT*=1").
   b. Find the row corresponding to the 10th period, and divide the "number of events" that have occurred by the size of the risk set for this period. The result is the (conditional) hazard probability for the 10th period, and can be checked against the percentage value in the output and the sample hazard estimate in the first panel of Table 11.1, also for the tenth period.
   c. Use the "cumulating" formula for survival probability in Equation (10.4) on p. 335 of *ALDA* to estimate the sample survival probability in: (a) period 7 (remember that everybody begins by "surviving" in grade 6), and (b) period 8. Check your values against the appropriate entry in Table 11.1.
6. Now, fit your first discrete-time hazard model to the *firstsex* data. This first model will replicate Model A in Table 11.3, on page 386 of *ALDA*.
   a. Begin, by "clearing all" from the *Editor* and the *Log*, and deleting the output files from the *SAS Explorer* window if you wish. Then, type the following new program into the *SAS Editor*:

```
proc logistic data="c:/????/firstsex_pp.sas7bdat" descending;
  model event=D7-D12 / noint;
run;
```

      i. You could also simply add the new code to the end of the previous code, if you want, inserting it immediately *before* the "run;" command – if you do this, then *both* the cross-tabulation and the logistic regression analysis will execute sequentially when you run the new program.
      ii. In *SAS*, logistic regression analyses are conducted using "proc logistic" in the person-period dataset. The hypothesized DTSA model is specified in the "model" statement, with the outcome variable ("*EVENT*") listed on the left of the "=" sign and the predictor variables (*D7* thru *D12*) listed on the

right of the "=" sign. The "noint" option in the "model" sentence eliminates the regular intercept, in favor of the multiple intercepts associated with the complete set of time dummies. The "descending" option in the "proc" sentence defines the value "1" as *event occurrence* for the *EVENT* variable. (*PROC LOGISTIC* treats its outcome as categorical, not numerical, and so it doesn't know until you tell it, that "1" is higher than "0." Try rerunning the program without the "descending" option, and check out what happens to the logistic regression coefficients!)

b. Execute the program in the usual way and debug until you successfully obtain output.

c. Print a copy of the output for your records and, on the output, identify and compare the following with entries in Table 11.3 on p. 386:

   i. The *parameter estimates* associated with time dummies, *D7* thru *D12*. Notice that their anti-logged versions are listed as "Odds Ratio Estimates" beneath the estimate listing.

   ii. The *standard errors* associated with time dummies, *D7* thru *D12*.

   iii. The *deviance statistic* associated with the model fit. It is listed as the entry for "-2 Log L" under "Model Fit Statistics, With Covariates".

   iv. The *AIC* statistic. It is listed under "Model Fit Statstics, With Covariates."

d. Using Equation (11.15) on p. 387, transform the coefficient associated with the *D12* predictor to obtain a fitted hazard probability and compare it to the value contained in the last row of Table 11.4 on p. 388.

7. Now, refit the discrete-time hazard model with the *PT* and *PAS* predictors included, in order to replicate Model D in Table 11.3 (p. 386).

   a. Begin, by clearing the appropriate *SAS* windows and editing the existing program in the *SAS Editor*, to read:

```
proc logistic data="c:/????/firstsex_pp.sas7bdat" descending;
  model event=D7-D12 pt pas / noint;
run;
```

      i. In *SAS*, multiple predictors like *PT* and *PAS* can be added to a model statement by listing them out on the right hand side of the "=" sign, separated by one or more spaces.

   b. Execute the program in the usual way.

   c. Print a copy of the output for your records and on it, identify and compare the following with entries in Table 11.3 on p. 386, under Model D:

      i. The *parameter estimates* associated with the *PT* and *PAS* predictors.

      ii. The *deviance statistic* associated with the model fit.

d. Compare your "odds ratio estimate" obtained for the *PT* predictor with the value given in text on p. 389, beneath Equation (11.16). Notice that your value differs from the text value (1.94 vs. 2.40) because the text version is obtained in Model B, which estimates the impact of *PT* in the absence of a control for *PAS*. Your estimate controls for the main effect of *PAS*.

8. In Chapter 12 of *ALDA*, we describe how to replace the completely general specification of *TIME* in the hazard model with more parsimonious polynomial specifications and for comparing the fits of these models in order to determine which specification is most effective. This approach can also be adopted with the *firstsex* example, even though it is probably not necessary, given that *TIME* contains a small number of discrete time-periods.

   a. Begin, by clearing the appropriate *SAS* windows and editing the existing program in the *SAS Editor*, to read:

```
data sex;
  set "c:/????/firstsex_pp.sas7bdat";
  one=1;
  t1=period-7;
  t2=t1*t1;
proc logistic data=sex descending;
  model event = one pt pas / noint;
proc logistic data=sex descending;
  model event = one t1 pt pas / noint;
proc logistic data=sex descending;
  model event = one t1 t2 pt pas / noint;
run;
```

      i. In this program, you are asked to implement a different strategy for entering the data into the analysis, using a "data" step to create a temporary SAS dataset that exists only within this particular run. We have called this temporary dataset, "sex." The *firstsex_pp* system datafile is read into the temporary dataset by the "set" command on the second line.

      ii. We have adopted this new approach to data entry so that you can easily create new polynomial representations of time from the "*PERIOD*" variable already present in the dataset. The new representations are: (a) the variable *one*, which is constant and equal to 1 in every record (and will serve as the new intercept in the polynomial models, see Table 12.1, p. 411), (b) the variable *t1*, which represents linear time (re-centered on grade 7), and (c) the variable *t2*, which represents quadratic time.

      iii. The program also contains multiple "proc logistic" paragraphs, each of which requests that a different discrete-time hazard model be fit. Notice that we ask each of these implementations of the procedure to fit the required model using the temporary dataset "sex" and not the original dataset, in order to make use of the newly minted polynomial representations of time.

b. Execute the program in the usual way.

c. Print a copy of the output for your records and, on the output, identify the *deviance* and *AIC statistics* associated with each model fit. Compare these fits to the deviance statistic for the completely general model (containing both the *PT* and *PAS* predictors) obtained earlier. Notice, as argued in our presentation, that the deviance statistics for the *constant model* and the *completely general model* "bracket" the deviance statistics for all the polynomial specifications.

d. Using the strategies of fit comparison described in our presentation and book, determine which polynomial representation best represents the shape of the discrete-time hazard function.

e. Use the "odds ratio estimates" on the output to interpret the main of effects of the *PT* and *PAS* predictors, in your selected model.

9. In Chapter 12 of *ALDA*, we also describe how to test the adequacy of the proportional odds assumption made in many discrete-time hazard models, by including interactions between *TIME* and substantive predictors. This approach can also be adopted with the *firstsex* example.

    a. You will test for the failure of the proportionality assumption with regards the *PT* predictor in a model that uses a linear specification of *TIME*, investigated above.

    b. Begin by editing your existing program in the *SAS Editor* to:

        i. Create an interaction between the *PT* predictor and linear time, *t1*, in the "data" step, give it a variable name of your own choosing.

        ii. Fit a discrete-time hazard model that contains the predictors *ONE*, linear time, *PT*, and the new interaction.

    c. Execute the program in the usual way.

    d. Print a copy of the output for your records and, on the output, identify the statistic that confirms or disconfirms the adequacy of the proportionality assumption. What is your conclusion?

10. If you have any time left, take a look at the programs and output for *ALDA* Chapters 11 and 12 on the *UCLA* support site or work on getting your own data in shape for your subsequent analyses on Friday morning.