Data-Analysis Exercise

Treating TIME More Flexibly/Discontinuous & Non-Linear Change

(ALDA, Chapters 5 & 6, pp. 138-242)

Purpose of the Exercise

This data-analytic exercise supplements our presentation that discussed ways of treating *TIME* more flexibly and investigating discontinuous and non-linear change with the multilevel model for change. The immediate purpose for the exercise is to supplement and extend your implementation of multilevel modeling with *SAS PROC MIXED* and *NLMIXED*.

General Resources for Conducting the Exercise

As before, while we provide self-contained instructions below for conducting the 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/ .

Datasets

The datasets used in this exercise were formatted by *UCLA Academic Technology Services* and are available for download, as text and system files from their website. For our current purposes, the datasets have 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 through the requested tasks:

1. In this exercise, you will use the data on *hourly wages and GED receipt among high-school dropouts* featured in *ALDA* Table 6.1, pp. 192 ff., and in our presentation. Please have this exhibit close at hand.
2. The files that contain these data have the prefix *wages*. They appear in the data directory several times with different suffixes denoting different versions of the file -- as raw text and as system files appropriate for analyses with the different software packages.
    a. Examine the *raw text person-period dataset*, called *wages_pp.txt*, by opening *Windows Explorer* and clicking on the filename in the appropriate data directory on your hard drive. Windows *Wordpad* should then display the contents of the file (it is too large for *Notepad*). Make sure that the first four lines of the dataset match the following (line-1 wraps onto a second line):

```
"ID","LNW","EXPER","GED","POSTEXP","BLACK","HISPANIC","HGC","HGC_9","UERATE","U
E_7","UE_CENTERT1","UE_MEAN","UE_PERSON_CENTERED","UE1"
31,1.491,.015,1,.015,0,1,8,-1,3.215,-3.785,0,3.215,0,3.215
31,1.433,.715,1,.715,0,1,8,-1,3.215,-3.785,0,3.215,0,3.215
31,1.469,1.734,1,1.734,0,1,8,-1,3.215,-3.785,0,3.215,0,3.215
```

b. Notice that the raw text person-period data-file contains comma-delimited text, with the first line containing the variable names, in quotes: *ID*, *LNW*, *EXPER*, *GED*, *POSTEXP*, *BLACK*, *HISPANIC*, *HGC*, *HGC_9*, *UERATE*, *UE_7*, *UE_CENTERT1*, *UE_MEAN*, *UE_PERSON_CENTERED*, *UE1*. For the moment, we focus only on *ID* thru *UE_7*. By consulting *ALDA*, match these variable names with the variable descriptions on pp. 146-151 and 191-194. Make sure that you can line up the names of the variables with their values for the first case, across all the three waves of data, and that you understand why they take on the values that they do:

   i. Notice that variables *HGC_9* and *UE_7* are re-centered versions of the original variables, obtained by subtracting 9 (years) and 7 (percentage points) from the variables measuring the dropout's highest grade completed and local unemployment rate, respectively. The re-centered versions are used in subsequent analyses, as described in *ALDA*.

   ii. Notice, as described in *ALDA* and in the presentation, that this is an unbalanced dataset with varying numbers of waves of data and varying spacings of those waves across individuals. In addition, several of the predictor variables -- *EXPER*, *GED*, *POSTEXP*, *UERATE* -- are *time-varying*.

3. Boot up *PC-SAS* on your workstation.
4. Begin your analyses of the *wages* dataset using *SAS*.
   a. Read the *wages* person-period system file (*wages_pp.sas7dat*) into *SAS* and display the empirical log-wage trajectory for dropout #2365 (who is also featured in Table 6.1), using the following *SAS* code (replacing the data file name with a name appropriate for your system):

```
proc plot data="C:/????/wages_pp.sas7bdat";
  where id=2365;
  plot lnw*exper / vaxis=0 to 3 by 1 haxis=0 to 14 by 1;
run;
```

   b. Execute your program, obtain output and print it out for your records.

i. Given that this dropout earned his GED between the fourth and fifth waves of data collection, inspect the observed log-wage trajectory for evidence of a potential discontinuity on receipt of the GED? Does it impact the elevation, the rate of change, both, neither?

5. Fit a sequence of multilevel models for change, mirroring Table 6.2, where a variety of fixed effects and variance components have been added and removed from a baseline model (Model A) in order to reach a reasonable final model (Model F).

   a. First, work with baseline Model A. The L1/L2 specification of this model is given in Equation (6.6) on p. 201 of *ALDA*. On a separate sheet of paper, perform the required substitutions in order to obtain the corresponding composite specification (which is not listed in *ALDA* and so is not available for checking). According to the first row of Table 6.2, however, it must contain 5 fixed effects and 4 variance components. You will need the composite model in writing your *SAS* code below.

   b. Now fit Model A to the *wages* data.

      i. Begin, by clearing the *Editor* and the *Log*, and then type the following new program code into the SAS *Editor*:

```
proc mixed data='c:\????\wages_pp' method=ml;
  title1 'Model A: EXPER, HGC-9, BLACK*EXPER, UE-7';
  class id;
  model lnw=exper hgc_9 exper*black ue_7 / solution notest;
  random intercept exper / subject=id type=un;
run;
```

      ii. Examine the correspondence between the model statement in this code, and the composite model you have created. Notice that much of the program remains that same as previous code that you have written for *PROC MIXED*, but that the "model" and "random" sentences reflect the presence of the five fixed effects (the intercept, plus the effects of the four predictors *EXPER*, *HGC_9*, *EXPER*BLACK*, *UE_7*) and four random effects (σ_ε^2, σ_0^2, σ_{10} and σ_1^2) that you needed to include in the model.

      iii. Execute the program in the usual way and obtain output. The program may take some time to run as it is complex and there are a large number of cases (our execution took about 25 seconds of CPU time).

         1. Print a copy of the output for your records and, on the output, identify the usual parameter estimates.

2. Compare the *deviance statistic* associated with the model fit, listed as the entry for "-2 Log Likelihood" under "Fit Statistics" with the entry in Table 6.2.

c. Now, modify your *SAS* code to fit the multilevel model for change *five more times*, following the taxonomy of models given in Table 6.2 from B through F. All models can be fitted in the same run by copying and pasting the basic code five more times into the program (before the "run;" command) and editing the "model" and "random" statements in sensible ways according to the specifications in Table 6.2. (If you cannot figure out the code, take a look at the *UCLA* support site, under "Chapter 6, Table 6.2.")

d. Execute the program in the usual way.

e. Print an edited copy of the output for your records (delete the pages with the endless streams of ID #'s) and:

i. Compare the obtained deviance statistics to those listed for Models B thru F in Table 6.2.

ii. Compare the parameter estimates for Model F to those presented in Table 6.3.

iii. Using the parameter estimates for Model F, draw a sketch graph, by hand, of the fitted log-wage trajectory of a prototypical Black dropout with 9 years of high school completed who obtains the GED in his third year on the job in a district with an employment rate of 7%. Compare it to the appropriate fitted trajectory in Figure 6.3.

6. As a final part of this data-exercise, here's something on the new *SAS* procedure, *PROC NLMIXED*, for fitting truly non-linear multilevel models. Here, we will use it with the Fox n'Geese data, contained in the *foxngeese* set of data files.

a. The nonlinear multilevel model that you will fit is the restricted logistic model in Equation (6.8) on p. 228 of *ALDA*.

b. The *SAS* code (don't forget to change the filename) is:

```
proc nlmixed data="c:/????/foxngeese_pp.sas7bdat" maxiter=1000;
  title1 'Model A: Unconditional logistic growth trajectory';
  parms G00=12 G10=.1 s2e=20 s2u0=.3 cu10=0 s2u1=.01;
  model NMOVES ~ normal(1+19/(1 + G00*(exp(-(G10*GAME +u0 +u1*GAME)))),s2e);
  random u0 u1 ~ normal([0,0],[s2u0,cu10,s2u1]) subject=id;
run;
```

c. The *PROC NLMIXED* routine is a very general, powerful and flexible routine that you can use to fit pretty much any kind of nonlinear multilevel model that you want, assuming all kinds of error distributional properties. It has many options that are not illustrated in the box above, and it is certainly worth an expedition to the *SAS* website to find out more about it. Briefly:

i. In the "proc nlmixed" line, the "maxiter" option is used to put an upper limit on the number of iterations that the routine will perform before stopping (unless it converges first).

ii. The "parms" command establishes starting values for the parameter estimates in the iterative estimation process. You can pretty much pick any values that you want that are sensible (you can't choose negative values for the starting values of variance components, for instance). There are two kinds of parameters, fixed and random, matching the content of Equation (6.8) and (6.9) – compare the code to the equations.

iii. The "model" statement is slightly different from its *PROC MIXED* peer. In *NLMIXED*, you specify the "model" along with a choice of distribution for the outcome, *NMOVES* – here, "normal( )." Inside the parentheses associated with the distribution statement, you write out the hypothesized model algebraically, making up names for each of the parameters, and you include a symbol to represent the level-1 error variance (here, s2e).

iv. The "random" statement permits you to specify the level-2 error covariance structure, using "u's" instead of "ζ's" (as in Equation (6.9)). Here, we assume that the two level-2 errors have the usual error structure and are normally distributed.

v. The "subject" command is as it was before.

d. Execute the program and obtain output. Compare the output to the estimates for Model A in Table 6.6 of *ALDA*. Isn't it neat? The world is your non-linear oyster.

7. If you have any time left, take a look at the programs and output for *ALDA* Chapters 5 and 6 on the *UCLA* support site, or work on getting your own data in shape for your subsequent analyses on Friday morning.