

The use of PROC MIXED for Analyzing Cohort-Sequential Designs

C. Nathan Marti, The University of Texas, Austin, TX

Introduction

Cohort-sequential designs offer an efficient method for measuring change across time. Longitudinal designs are useful for tracking change in the same individual, however they require substantial amounts of time to complete the study. Cross-sectional designs allow researchers to compare several age groups at the same time and thus measure change across time, but as this is a between-subjects comparison, it is not possible to measure individual change. Both longitudinal and cross-sectional designs suffer from potential cohort effects: longitudinal designs are limited to a single cohort that may have unique characteristics relevant to the period in time in which they are being studied; cross sectional designs potentially confound age differences with cohort differences where cohort differences may be a result of a unique environmental factor that only affects a particular age group. Cross-sequential designs measure different age groups across multiple time points. By doing so, a change across time can be analyzed for a much greater time period than is required to complete a study. For instance, the example used in the present paper follow children that are age nine, ten, eleven, and twelve at wave one of a study for three years, thus measuring change across time for a seven year period in a span of three years.

A single dataset (Hetherington, Clingempeel, Anderson, Deal, Lindner, & Stanley-Hagan, 1992) is used to illustrate the use of PROC MIXED for cohort-sequential designs. The study examined children's involvement in mutual activities with their mothers between the years of 1984 and 1986. Children ranged from nine to thirteen years of age at the onset of the study and were measured at three yearly intervals.

While there are several advantages of cohort-sequential designs, they present difficulties for ordinary least squares (OLS) analyses. Longitudinal designs can be analyzed as repeated measurements with time points treated as within-subjects factors and cross-sectional designs can be analyzed as a factorial analysis of variance where groups at different ages are treated as between-subject factors. However, neither of these approaches is appropriate for the analysis of cohort-sequential designs as there are potentially both within and between-subjects factors, but no subject has complete data at all levels. For instance in the present example, responses are measured at seven different ages, but each individual is only measured on

three occasions and therefore has missing data for the other four time points. In fact, some comparisons confound between and within-subjects comparisons: take for instance a comparison between nine and twelve-year-olds in a study with three yearly waves. In this comparison, the subjects that began the study as nine-year-olds would be compared with twelve-year-olds, but the participants that began the study as ten-year-olds would be compared with themselves at twelve years of age and children from other cohorts at twelve years of age. Thus, the line between within and between-subjects comparisons is blurred in the cohort-sequential designs.

Historically, researchers have analyzed cohort sequential designs longitudinally (Anderson, 1995). This is likely a result of the fact that cases with missing data are dropped in OLS analyses and, as all cases in a cohort-sequential design have missing data if it were assumed that they have values for all time points, all cases would be dropped. To analyze cohort-sequential designs longitudinally, researchers essentially collapse all age groups within each wave of the study. For example, in the example data used here, this would be a comparison between three waves in which the average age at wave one is eleven years of age, twelve years of age at wave two, and thirteen years of age at time three. Considering the range of ages at each of these time points, it is apparent that there is a large amount of information that is being lost by this comparison. For example, at wave one, the range of ages is between nine and thirteen years of age. Given the disadvantages of OLS approaches, it is prudent to consider what a more ideal form of analysis may be. PROC MIXED ability to handle missing data makes it an ideal procedure for analyzing planned patterns of missing data as will be described in this paper.

Preparing Data for Analysis Using PROC MIXED

Using PROC MIXED to analyze cohort-sequential designs is largely a result of creating an appropriately structured dataset. Before using PROC MIXED there are some common general issues to consider and as well as issues specific to analyzing cohort-sequential designs. A common general consideration is that datasets are often organized in a multivariate format such that there is a single line for each case and a column for each data point. PROC MIXED requires that data be organized in a univariate format so that there is a single row for each measurement occasion. In addition to the general dataset requirements of PROC MIXED, cohort-sequential designs require that there is a variable for every possible

point in time that is being examined. Of course, in a cohort-sequential design, the number of responses per subject is less than or equal to the number of waves in the study. For instance, the present example has seven age groups ranging from nine to fifteen years of age that were measured at three time points. Thus, a participant in the study who was eleven years of age at the beginning of the study would have data points for age eleven, twelve, and thirteen, but would have missing data for ages nine, ten, fourteen, and fifteen.

The present example first considers the cohort-sequential issue of creating response variables for every possible measurement occasion. First consider the organization of the data in a typical multivariate format as shown below. The data shown here contains an identification variable, *famid*, a cohort variable, *age*, that represents the age of the participant at the beginning of the study, and responses for each time point, *eaf1*, *eaf2*, and *eaf3*.

	FAMID	AGE	EAF1	EAF2	EAF3
1	01	9.00	54.00	52.00	59.00
2	010	12.00	60.00	64.00	60.00
3	011	11.00	57.00	54.00	75.00
4	012	10.00	46.00	54.00	55.00
5	013	13.00	43.00	47.00	62.00
6	014	11.00	53.00	42.00	47.00
7	017	10.00	79.00	56.00	42.00
8	018	9.00	36.00	71.00	74.00
9	019	11.00	50.00	53.00	46.00
10	02	11.00	43.00	68.00	55.00
11	020	11.00	50.00	55.00	55.00

To create variables for each possible age, a series of IF-THEN statements are used. In the example below, there is a separate IF-THEN statement for each of the cohorts. The statement is used to assign each of the dependent measures to the proper wave.

```
DATA mixed.two;
  SET mixed.one;
  IF age = 9 then do;
    Age09 = eaf1; age10 = eaf2; age11 = eaf3;
  END;
  IF age = 10 then do;
    age10 = eaf1; age11 = eaf2; age12 = eaf3;
  END;
  IF age = 11 then do;
    age11 = eaf1; age12 = eaf2; age13 = eaf3;
  END;
  IF age = 12 then do;
    age12 = eaf1; age13 = eaf2; age14 = eaf3;
  END;
  IF age = 13 then do;
    age13 = eaf1; age14 = eaf2; age15 = eaf3;
```

```

END;
DROP age eaf1 eaf2 eaf3;
RUN;

```

As the syntax shown above, seven variables are created, *age09*, *age10*, *age11*, *age12*, *age13*, *age14*, and *age15*. The values of these variables are assigned with regard to the cohort to which a participant belongs. For example, a participant who was nine years of age at the beginning of the study has a value of 9 for the variable *age*, and therefore the first IF-THEN statement is used to calculate the dependent variables. Thus, such a participant would be assigned the value of the dependent measure on the first wave for the *age09* variable, the value of the second wave for the *age10* variable, and the value of the third wave for the *age11* variable. Values for *age12*, *age13*, *age14*, and *age15* are missing as a participant who was nine years of age at the beginning of the study would be age eleven at the end of the study and therefore would not have data points for ages twelve through fifteen. The dataset with the new variables calculated is shown below.

	FAMID	age09	age10	age11	age12	age13	age14	age15
1	01	54	52	59	.	.	.	.
2	010	.	.	.	60	64	60	.
3	011	.	.	57	54	75	.	.
4	012	.	46	54	55	.	.	.
5	013	.	.	.	.	43	47	62
6	014	.	.	53	42	47	.	.
7	017	.	79	56	42	.	.	.
8	018	36	71	74	.	.	.	.
9	019	.	.	50	53	46	.	.
10	02	.	.	43	68	55	.	.
11	020	.	.	50	55	55	.	.
12	021	.	.	74	57	70	.	.
13	022	66	30	34	.	.	.	.
14	023	44	51	70	.	.	.	.
15	025	.	.	.	71	58	62	.

As can be seen in the dataset above, each participant has three responses, the first of which corresponds with their age at the first wave of the study. For example, compare the first case with the previous dataset. This case was nine years of age when the study began and therefore has values for the *age09*, *age10*, and *age11* variables.

Following the creation of the variables representing all possible ages in the study, the next step in preparing the data is to transpose the data from a multivariate to univariate format. To do this, we use

PROC TRANSPOSE in the present example. The syntax below illustrates the use of PROC TRANSPOSE to convert the data into a univariate dataset.

```
PROC TRANSPOSE DATA=mixed.two OUT=mixed.two NAME=age PREFIX=score_;
  VAR age09-age15;
  BY famid;
RUN ;
```

In the PROC TRANSPOSE above, the variables *age09* through *age15* are transposed so that they are in a single column. The NAME argument creates a new variable named *age* that stores the name of the variable being transposed, which in this case is the variables *age09* through *age15*. The PREFIX argument provides a prefix for the name of transposed variables. Thus, the dataset that is created contains the identification variable, *famid*, the new variable *age*, which contains the names of the transposed variables as values, and *score_1*, the variable containing the values of the scores. The dataset used in the present example is shown below:

	FAMID	age	score_1
1	01	age09	54
2	01	age10	52
3	01	age11	59
4	01	age12	.
5	01	age13	.
6	01	age14	.
7	01	age15	.
8	010	age09	.
9	010	age10	.
10	010	age11	.
11	010	age12	60
12	010	age13	64
13	010	age14	60
14	010	age15	.
15	011	age09	.
16	011	age10	.
17	011	age11	57
18	011	age12	54
19	011	age13	75
20	011	age14	.
21	011	age15	.
22	012	age09	.
23	012	age10	46
24	012	age11	54
25	012	age12	55
26	012	age13	.

The data is now ready to be analyzed using PROC MIXED. In the above dataset, the independent variable, *age*, is a categorical variable that enables you to compare mean values at different ages. This is described in the section, *Comparisons Between Time Points*, below. Additionally, you may want to construct a random coefficient model in which individuals' intercepts and slopes are allowed to vary and this is illustrated in the section, *Individual Growth Curve Models in a Cohort-Sequential Design Using PROC MIXED*.

Comparisons Between Time Points Using PROC MIXED

One possible analysis of data from a cohort-sequential design is to compare values of the dependent variable at different time points. To do this, the independent variable should be defined as a string variable. The syntax for a model that tests for a main effect of the variable *age* is shown below.

```
PROC MIXED DATA = Mixed.two;  
  CLASS famid age;  
  MODEL score_1 = age;  
  REPEATED age / SUBJECT = famid TYPE = un;  
RUN;
```

Categorical variables are listed in the CLASS statement. In the present example, this includes the identification variable, *famid*, and the variable representing individuals' ages, *age*. The MODEL statement describes the model being tested where the dependent variable is on the left side of the equal sign and the independent variable or variables are on the right side. The model being tested in the present example is using *age* to predict participants' score. The REPEATED statement indicates that *age* is a within-subjects variable. The subject unit is defined by the SUBJECT option, indicating that all responses with identical values of *famid* are from the same respondent. The TYPE option specifies the covariance structure; in this case, the unstructured covariance option is selected. The principle analysis of interest from the model shown above is to test for a main effect of age. This test is obtained in the *Type 3 Tests of Fixed Effects* output.

Type 3 Tests of Fixed Effects				
	Num	Den		
Effect	DF	DF	F Value	Pr > F
age	6	158	1.07	0.3850

In the present example, the small *F* value indicates that there is not a main effect of age. Although, there is not a main effect of age, examination of the data may indicate that there are particular ages that are different from the others. In the present example, ages nine through twelve all have similar ages, whereas the older ages show a decline in their value in the dependent variable. To analyze specific comparisons between ages, you might consider using the CONTRAST statement to construct custom hypothesis tests. The following example illustrates two uses of the contrast statement: the first compares twelve-year-olds with fourteen-year-olds, and the second compares nine through twelve-year-olds with fourteen-year-olds.

```
PROC MIXED DATA = Mixed.two;
  CLASS famid age;
  MODEL score_1 = age;
  REPEATED age / SUBJECT = famid TYPE = un;
  CONTRAST '12 versus 14' age 0 0 0 1 0 -1 0;
  CONTRAST '9-12 versus 14' age 1 1 1 1 0 -4 0;
RUN;
```

The custom contrasts can be examined in the *Contrasts* table in the output.

Contrasts				
	Num	Den		
Label	DF	DF	F Value	Pr > F
12 versus 14	1	158	2.44	0.1205
9-12 versus 14	1	158	4.07	0.0453

This table lists each contrast separately. The first contrast, *12 versus 14*, compares mean values of twelve and fourteen-year-olds, and shows that there is not a significant difference between the two age groups. The second comparison, *9-12 versus 14*, between the fourteen-year-olds and the nine through twelve-year-olds

indicates that fourteen-year-olds have a significant difference in their scores than the four ages with which they are being compared.

Individual Growth Curve Models in a Cohort-Sequential Design

Individual growth curve models allow researcher to explicitly model individual growth and present many advantages to repeated measures analyses (Bryk, & Raudenbush, 1992). To do so, a multilevel model is constructed in which time points are level-1 units and individual are level-2 units. Thus, time points are nested within individuals. By constructing such a model, you can first examine the hypothesis about whether it is appropriate to use a single regression model for all subjects in you dataset. If you have a significant effect for level-2 error terms, it not appropriate to model your data as a regression equation as a single intercept and slope are not sufficient for an individual who vary on these parameters. One advantage of using a multilevel model is that it includes error terms form both level, and therefore, the effects for variables are not potentially confounded with the variances due to individual's variation.

The present example also illustrates the use of a continuous predictor variable, which provides output resembling a regression analysis. A critical difference between the previous example and regression analyses is that when age is treated as a categorical variable, PROC MIXED makes contrast comparisons between ages or between a group of ages and other ages as seen in the contrast example above. In contrast, when the predictor is continuous, PROC MIXED measures the change in the dependent variable that can be attributed to each unit of the independent variable. In the present example, the change would be the amount of increase or decrease in the scores measuring children's involvement that can be accounted for by their age.

Prior to using proc mixed to perform a regression style analysis, the independent variable in the example dataset would need to be converted to a continuous variable. This is done using the following DATA step in which a new, numeric-formatted variable, *age_2* is created using the SELECT statement.

```
DATA mixed.three;  
  SET mixed.two;  
  SELECT (age);  
    WHEN('age09') age_2 = 9;  
    WHEN('age10') age_2 = 10;  
    WHEN('age11') age_2 = 11;  
    WHEN('age12') age_2 = 12;  
    WHEN('age13') age_2 = 13;
```

```

        WHEN('age14') age_2 = 14;
        WHEN('age15') age_2 = 15;
        OTHERWISE;
        END;
RUN;

```

There are some important options introduced in the PROC MIXED example below. First, note that the COVTEST option has been added to the procedure statement. This requests covariance parameter estimates and their associated test statistics. Next, note the RANDOM statement. This statement is used to define random terms in the model. Here, the intercept, INT, and the slope parameter for the variable, *age_2* are defined as random terms. This indicates that these terms vary randomly across subjects. The TYPE statement defines the covariance structure in the same manner as the above example and the subject unit is defined by the SUBJECT statement indicating that *famid* is the level-2 unit in this model.

```

PROC MIXED DATA = Mixed.two COVTEST;
    CLASS famid age;
    MODEL score_1 = age_2 / SOLUTION ;
    RANDOM INT age_2 / TYPE = UN SUBJECT = famid;
RUN ;

```

To examine the unique effects of individuals, look at the Covariance Parameter Estimates table. In the present example, the table shows that none of the random effects were significant, indicating that they are not a necessary component of the model. That is, all of the *z* values are small and their associated *p* values are larger than .05. The first parameter listed is the intercept for which there is not a significant difference across subjects (*p* = .14), and the third term is the slope, which again does not differ across subjects (*p* = .13).

Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	FAMID	607.54	562.51	1.08	0.1401
UN(2,1)	FAMID	-48.5677	47.1352	-1.03	0.3028
UN(2,2)	FAMID	4.3959	3.9571	1.11	0.1333
Residual		125.37	10.9282	11.47	<.0001

The *Solution for Fixed Effects* table shown below contains information about the effect of age on individuals' scores. The coefficients can be interpreted as a standard regression equation as there are not level-2 covariates. Examining the table, it can be seen that there was a significant effect for *age* in this model as the *p* value associated with the *t* value is smaller than .05.

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	T Value	Pr > t
Intercept	64.2621	5.6479	158	11.38	<.0001
age_2	-0.9808	0.4728	158	-2.07	0.0397

Conclusions

The ability of PROC MIXED to handle missing data makes it an ideal procedure to analyze cohort-sequential designs which present analytic difficulties to OLS methods by forcing analysts into using between or within-subjects comparisons. By constructing a planned pattern of missing data and treating responses as repeated measurements, PROC MIXED can be used to analyze cohort-sequential designs. While the present discussion has focused on cohort-sequential designs, it also has applications to other

designs that employ repeated measurement with missing data. Most notably, it can easily be applied to longitudinal designs in which there is missing data as a result of participants failing to participate in all waves of a study or due to dropouts. Deleting these cases from analyses could bias results, as the participants that fail to participate in all measurement occasions of a study are likely to be different than those that do complete all phases of the study. Thus, employing the approach used in the present paper could potentially serve to improve analyses of longitudinal data in addition to the improvements already discussed.

References

- Anderson, E. R. (1995). Accelerating and maximizing information from short-term longitudinal research. In J. M. Gottman (Ed.), *The Analysis of Change* (pp. 139-163). Mahwah, N.J.: Lawrence Erlbaum Associates.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical Linear Models*. Newbury Park, CA: Sage.
- Hetherington, E. M., Clingempeel, W. G., Anderson, E. R., Deal, J. E., Lindner, M. S., & Stanley-Hagan, M. (1992). Coping with marital transitions: A family systems perspective. *Monographs of the Society for Research in Child Development*, 57 (2-3, Serial No. 227).
- Littell, R.C., Milliken, G.A., Stroup, W.W., & Wolfinger, R.D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute, Inc.
- Singer, J. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 24, 323-355.