Repeated Measures ANCOVA with the MIXED and GLM Procedures: 
Examining an Intervention to Reduce Childhood Obesity
Lauren Cook, University of Southern California, Los Angeles, CA
Jaimie N Davis, RD, PhD, University of Texas, Austin, TX
Donna Spruijt-Metz, MFA, PhD, University of Southern California, Los Angeles, CA
Nicole Gatto, MPH, PhD, Loma Linda University, Loma Linda, CA

ABSTRACT
This paper describes a basic use of the MIXED procedure in SAS® to conduct a repeated-measures Analysis of Covariance (ANCOVA), building on preliminary analyses from PROC GLM. The data were collected at two separate occasions, and include a dichotomous variable that serves as the main predictor, several dependent variables of interest, and covariates that are predominantly time-independent, as well as some that are time-dependent. The specific goal is to examine the effect of a nutrition and gardening intervention with elementary school children in Los Angeles, as compared to a control group. The outcomes examined are childhood obesity, measured by body mass index, and dietary intake, specifically vegetable consumption.

The data are initially structured at the person level with each observation containing individual data from both measurement occasions. The main outcome of interest (child body mass index) requires the use of time-independent covariates only, and thus is examined first using PROC GLM.

Analysis of the secondary outcome of interest (vegetable intake) includes daily caloric intake as a time-dependent covariate, requiring the use of PROC MIXED. We use a data step array to restructure the data into a person-period format, where each observation corresponds to individual data from one measurement occasion only. The correct execution of the array was confirmed by repeating the GLM analysis of body mass index using MIXED. Finally, we examine the effect of the intervention on vegetable intake, controlling for both time-dependent and time-independent covariates, in MIXED.

This example illustrates repeated measures analysis using a simple research design suitable for users familiar with repeated measures studies but new to the MIXED procedure.

INTRODUCTION
This paper presents a basic example of PROC MIXED to perform a repeated measures Analysis of Covariance (ANCOVA), suitable for users already familiar with this technique in PROC GLM. Data are from a randomized controlled trial with participants exposed to one of two conditions, collected at baseline and follow-up. Two outcomes of interest are examined here. The first outcome, childhood obesity as measured by body mass index, has only time-independent predictors, which we examine first using GLM. The second outcome, vegetable consumption, necessitates the use of both time-dependent and time-independent covariates, thus a multilevel model is used in MIXED.

The intervention under study is the LA Sprouts nutrition, cooking and gardening program for obesity prevention in Los Angeles elementary school children. Four schools were randomized to either the intervention group or control (with delayed intervention), and clinical and dietary data were collected on all participating students at baseline and 12-week follow-up. The primary aims of this intervention are to decrease obesity and improve dietary behavior. The outcomes presented in this paper include body mass index (BMI, an estimate of obesity), and vegetable intake. Each outcome has several theoretical predictors that we use as a priori covariates. Our statistical question is to examine the controlled interaction of intervention group by time on our outcomes.

VARIABLES
The following variables are used in this analysis:

OUTCOME VARIABLES
BMI z-score: In children, BMI is adjusted for age and sex and compared to historically average values and expressed in terms of the number of standard deviations above or below the average.

Vegetable intake: Vegetable intake is measured via Block Kids Food Screener (last week version) and units are cup equivalents (CE). Vegetable intake is highly skewed and therefore log-transformed.
Repeated Measures ANCOVA with the MIXED and GLM procedures: Examining an intervention to reduce childhood obesity, continued

**PREDICTOR VARIABLES**

**Intervention Group**: Schools were randomized to one of two groups, intervention (LA Sprouts) or control.

**Wave**: Two schools (one control and one intervention) were evaluated during each wave of data collection. We cannot explicitly adjust for school site because Intervention Group is nested within this variable, such that each school only received one assignment. However, we use this variable to determine if there were any differences between schools in wave 1 versus wave 2. Although data are likely correlated within schools (i.e., children at one school are likely more similar to each other than to children at a different school), we do not use a multilevel model with school as a grouping variable because there are too few schools to create stable estimates.

**Mother Has Own Car**: This dichotomous variable is used as a proxy for socio-economic status. It is included as a covariate because a significant difference in mothers’ having a car was found at baseline between groups.

**Total Calories**: This variable was also determined via food screener; those with extreme values were excluded following cut points used with adult data.

**Age and Sex**: These variables are not included as covariates for BMI z-score, given the outcome is already adjusted. However, they are included in the dietary model.

All covariates above were determined a priori, and are retained in analytical models for theoretical purposes even when there is no evidence of statistical impact.

**DATA FORMAT**

Data are initially structured in a person-level format, with each observation consisting of all data for each study participant (that is, the observation includes both baseline and follow-up). This is an acceptable structure for general linear models, but is not suitable for a multi-level model allowing for time-dependent covariates. To reformat the data into a person-period format a data-step array is used; Table 1 shows the final format.

<table>
<thead>
<tr>
<th>ID</th>
<th>Intervention Group</th>
<th>Time</th>
<th>Sex</th>
<th>Vegetable (CE/d)</th>
<th>Energy (kcal/d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4.10</td>
<td>2239.3</td>
</tr>
<tr>
<td>101</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.80</td>
<td>2183.1</td>
</tr>
<tr>
<td>102</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1.04</td>
<td>1672.4</td>
</tr>
<tr>
<td>102</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.86</td>
<td>1251.7</td>
</tr>
</tbody>
</table>

Table 1: Example of Restructured Data

The code for data restructure is as follows:

```sas
data lasprouts_long;
  set lasprouts;
  age=pre_ageyrs;
  array Alog_veg(0:1) log_veg0 - log_veg1;
  array Aenergy(0:1) energy0 - energy1;
  array Abmiz(0:1) bmiz0 - bmiz1;
  do time = 0 to 1;
    log_veg = Alog_veg[time];
    energy = Aenergy[time];
    bmiz = Abmiz[time];
    output;
  end;
  keep id time car group wave sex age log_veg energy bmiz;
run;
```

This temporary database with a limited number of variables is created specifically for the research question(s) to be addressed. This approach of creating a limited database with only a subset of variables (as opposed to restructuring all data at once) is a helpful organizational strategy that aids to reduce programming and analysis errors. For subsequent analyses, univariate tests can be conducted in the original database as needed, and variables of interest can be restructured into a separate person-period database as needed.
STATISTICAL ANALYSES

BODY MASS INDEX

The first analytical step was to examine the main research question of whether or not exposure to the intervention had an impact on participant BMI. This was conducted using the original person-level database first, as follows:

```
proc glm data=lasprouts;
   class car group wave;
   model bmiz0 bmiz1 = group wave car;
   repeated time / printe;
   lsmeans group / stderr;
run;
```

The time by group interaction has an F value of 8.11 (p<0.01, Output 1), indicating that there is a significant effect of the intervention on participant BMI z-score following the 12-week program. Based on the output from the least squares means procedure, the intervention group decreased their average adjusted BMI z-score by 12.7% (0.11 absolute difference), whereas the control group decreased their average adjusted BMI z-score by 2.8% (0.03 absolute difference). We also find a significant time by wave interaction (F value = 8.11, p < 0.01), indicating that there is a difference between schools in wave 1 versus wave 2, such that students from wave two schools decreased their BMI z-score to a greater extent (0.11 absolute difference) than the students in wave 1 schools (absolute difference 0.03, output not shown).

![Output 1: Selected Output From BMI Z-Score Repeated Measures Analysis Using PROC GLM](image)

```
Output 1: Selected Output From BMI Z-Score Repeated Measures Analysis Using PROC GLM
```

Our next step is to conduct the same analytical test for BMI z-score using a multi-level model, to confirm the correct execution of the data restructuring. We use PROC MIXED as follows:

```
proc mixed data=lasprouts_long noclprint;
   class id car time group wave;
   model bmiz = time car group wave time*car time*group time*wave;
   repeated time / type=un sub=id ;
   lsmeans group*time ;
run;
```
Repeated Measures ANCOVA with the MIXED and GLM procedures: Examining an intervention to reduce childhood obesity, continued

In MIXED, for significance testing we do not need to include all categorical variables in the class statement, as we do with GLM, only those that are necessary grouping variables. These necessary variables differentiate subjects, time points, and intervention group (necessary for the ‘lsmeans’ statement, our main outcome); ‘car’ is not necessary, as this is a covariate that we just wish to adjust for. However, inclusion of non-grouping variables will marginally affect lease squares means estimates, so we choose to include these.

We chose the default method of estimation: restricted maximum likelihood, coupled with an unstructured covariance matrix (covariance structure indicated by the ‘type’ option). This estimation method (which can be changed using “method=”) provides variance component estimates equivalent to those in an ANOVA test when used with balanced data, as we have here.

For this test, the time by group interaction has an F value of 7.94 (p<0.01, Output 2). This value and adjusted means differ slightly based on the method of estimation and covariance structure selected, but we are satisfied that this test is consistent with results from the general linear model.

Output 2: Selected Output From BMI Z-Score Repeated Measures Analysis Using PROC MIXED

VEGETABLE INTAKE

The next research question is to examine the effect of the intervention program on vegetable intake in both groups. We again use PROC MIXED, adding additional a priori covariates to those used in the BMI analysis:

```
proc mixed data=lasprouts_long noclprint;
  class id car time group wave sex;
  model log_veg = time car group wave sex age energy time*car
    time*group time*wave sex*time age*time energy*time;
  repeated time / type=un sub=id ;
  lsmeans group*time ;
run;
```

This test indicates that the group by time interaction has an F value of 3.88 (p=0.05, Output 3), such that there is a marginally significant difference between the control and intervention groups following the 12-week program. Because the vegetable intake variable was log-transformed, we exponentiate the estimates in order to make comparisons. Both groups decreased their vegetable intake following the LA Sprouts program, yet the intervention group decreased their average adjusted intake by 6.1%, whereas the control decreased their average adjusted intake by 32.4%.
Repeated Measures ANCOVA with the MIXED and GLM procedures: Examining an intervention to reduce childhood obesity, continued

Output 3: Selected Output From Vegetable Intake Repeated Measures Analysis Using PROC MIXED

CONCLUSION

These analyses indicate that the LA Sprouts intervention is effective in decreasing child obesity, measured by BMI z-score, compared to controls. The significant difference between waves also indicates differing success of the program, possibly due to unmeasured school-level variables (perhaps teacher encouragement of program participation). The significant difference between groups may be attributable to differences in vegetable intake following the intervention (although not significant at p<0.05), yet it is surprising that both groups decreased their vegetable consumption. Seasonal variation and end-of-the-year parties may account for these unexpected dietary results, given that follow-up testing occurred during the final weeks before winter or summer vacation. It is also likely that other dietary or behavioral patterns contributed to difference in BMI between groups following the intervention.

This analysis provides a basic example of the MIXED Procedure for repeated measures analysis, using balanced data collected at only two time points. MIXED is a more flexible approach than the GLM Procedure because it allows for the use of time-dependent covariates, as we use here in the analysis of vegetable intake. Only two basic steps are required to use MIXED: the first is a data restructure (if needed, with which many users may already be familiar), and the second is execution and interpretation of the Procedure. Both Procedures are equally efficacious for repeated measures analysis, and choice between the two should be driven by the specific research question. GLM is more applicable to research questions pertaining to highly controlled environments (where individual characteristics do not change during the course of an experiment and measurement points are equally distributed), whereas MIXED allows for the analysis of unbalanced data or more complex models, as we see in this example.
REFERENCES


ACKNOWLEDGEMENTS

Lauren Cook would like to extend a huge thanks to David Pasta, her WUSS mentor. His feedback was invaluable.

This project is funded by NIH R21 DK094066-02.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Lauren Cook  
Enterprise: University of Southern California  
Address: 2250 Alcazar St, CSC-200  
City, State ZIP: Los Angeles, CA, 90089  
Work Phone: 323-442-2735  
Fax: 323-442-4103  
E-mail: laurenco@usc.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.