MWSUG 2018 – Paper HS-027

Tips, Tricks, and Traps on Longitudinal Data Analysis with Discrete and Continuous Times

Brandy R. Sinco, MS, University of Michigan, Ann Arbor, MI
Edith Kieffer, PhD, University of Michigan, Ann Arbor, MI
Michael Spencer, PhD, University of Michigan, Ann Arbor, MI
Gray Ficker, MDP, CHASS Center, Detroit, MI
Gretchen Piatt, PhD, University of Michigan, Ann Arbor, MI
Michele Heisler, MD, University of Michigan, Ann Arbor, MI

ABSTRACT

When longitudinal data are collected at discrete time points, such as at baseline, 6 and 12 months, compared to continuous times, both exploratory data analysis and linear mixed models need to be modified.

For data at discrete times, analysts can use Proc Corr to examine the correlation matrix by simply listing the variable names at each time point. In contrast, long datasets with continuous times must be transposed to a format that can be used with Proc Corr, by using the first and last functions. This presentation includes tips and tricks for viewing the empirical correlation structure when time is continuous.

When using Proc Mixed for a linear mixed model, some covariance structures differ between models with discrete and continuous times. SAS® offers covariance structures especially designed for continuous data, as well as structures that were designed for data with discrete times.

One trap and trick is the Estimate statement in Proc Mixed. For data at discrete times, time point coefficients are easily included in the Estimate statement. However, for polynomial models that contain time raised to various powers, the proper coding of time can make a difference between getting an "Inestimable error" versus a useful estimate.

This presentation features an example with diabetes intervention data collected over several years with a linear mixed model containing a third degree polynomial for time. When time was originally coded in months, the estimate statement in Proc Mixed produced an "Inestimable error". When time was re-coded in years, the estimate statement generated useful information.

Outline

1)_Introduction to Linear Mixed Model	Pages 2 - 3
2)_Empirical Correlation Matrices, Descriptive Statistics	Pages 3 - 6
3)_Exploratory Graphics	Pages 6 - 9
4)_Covariance Structures in SAS® Proc Mixed	Pages 9 - 12
5)_Linear Mixed Models with Discrete Time Points	Pages 13 - 14
6)_Linear Mixed Models when Time is Continuous	Pages 14 - 17
7)_Graphical Presentation of Results	Pages 17 - 19
8)_Conclusions	Page 19
9)_Acknowledgements	Page 19
10)_References	Page 20
11)_Contact Information	Page 20

1)_Introduction to the Linear Mixed Model (LMM).

First, recall the linear regression model with a simple random sample of size N.

- Let Y = outcome vector, dimension N rows × 1 column; aka dependent variable.
- Let X = matrix of m predictor variables, dimension N × (m+1), aka m independent variables + intercept.
- Let β_0 to β_m = Linear regression coefficients, where β_0 = intercept and β_j = mean increase in Y for a unit increase in the jth X variable, X_j. The β vector has dimensions (m+1) rows × 1 column.
- Let ε = Vector of error terms. $\varepsilon \sim N(0, \sigma^2)$; error terms assumed to have constant variance.

The linear regression equation will be $Y = X\beta + \varepsilon$ and the solution is found from least squares.

For example, let Y = height in feet and X_1 = age in years (X_1 10 to 17) and X_2 = gender (1=female, 0=male). For this model, assume that age is a fixed effect, meaning that age is being used to estimate a population average slope.

The model would be written as $Y = \beta_0 + \beta_1 age + \beta_2 gender + \varepsilon$.
If the solution to the regression were $\beta_0 = 4$, $\beta_1 = 0.1$, and $\beta_2 = -0.25$, this would increase that the population average rate of growth was 0.1 feet/year and that, on average, females were an average of ¼ foot shorter than males.

Linear Mixed Models (LMMs) are linear statistical models for continuous outcomes, in which the residuals are normally distributed, but not independent or not having constant variance[1]. The LMM is an expansion of the linear regression model. The X's and β's are the same as above, except that X's are designated as fixed effects. An additional matrix of random effects, Z, are added to the model, where Z_1 to Z_p are random effects with regression coefficients b_1 to b_p. The equations for the linear mixed model are:

- $Y = X\beta + Zb + \varepsilon$
- Instead of $\varepsilon \sim N(0, \sigma^2)$ in linear regression, $\varepsilon \sim N(0,\Sigma)$ because the residuals can be correlated in a LMM.
- $b \sim N(0, G)$, where G = covariance matrix of the random effects.
- Finally, b and ε are assumed to be independent.

In linear regression, ε is always modeled as a random effect with mean 0 and an individual estimate for each of Y_1 to Y_N. In the above example, if age were modeled as a random effect, the linear regression model would become a linear mixed model:

$Y = \beta_0 + \beta_1 age + \beta_2 gender + b_1 age + \varepsilon$.
This equation would produce a population average slope for age, in addition to an individual-level estimate for each person, analogous to the way that the random effect, ε, generates an error estimate for each person (or subject).

So, what is the difference between a fixed and a random effect? In 1960, Green and Tukey wrote[2] that "When a sample exhausts the population, the corresponding variable is fixed; when the sample is a small (i.e., negligible) part of the population the corresponding variable is random." So, if we have an experiment with a random sample of washing machines of 3 different brands and the outcome is reduction in dirt, we are interested in the effect of the brand of the washing machine, not the individual washing machine. So, the brand of washing machine is a fixed effect and the effect of the individual washing machine is a random effect.

Two guidelines for that I use for determining whether effects were fixed or random are:
First, consider clinic site as an example. If the experiment were repeated, would you choose the same clinics or is the individual clinic site just a random sample of many clinics? If the clinic site would be

chosen again, example eastside and southwest Detroit, then clinic site is fixed. On the other hand, if we select a random sample of clinics from a huge list, then clinic site would be a random effect. Second, if we are only interested in population-average effects, then the effect is fixed, such as age in the linear regression model. On the other hand, if we are interested in individual-level estimates, such as estimating pregnancy weight at the end of the first trimester for individual women, then gestational age would be a random effect.

2)_Empirical Correlation Matrices, Descriptive Statistics.

The first steps in longitudinal analysis are descriptive statistics and graphics. Means and variances at each time point need to be examined. If the variances are not equal between time points, choose covariance structure that allows for unequal variances. A guideline for standard deviations is whether any standard deviation is twice another standard deviation at a different time point, because this would correspond to an F statistic of 4 ($F = s_1^2/s_2^2$). In Table 2.1 (Proc Corr Summary Statistics), all variance ratios are <4, so heteroscedacity (unequal variances) probably isn't an issue.

For data with discrete times, such as baseline, 6 months, and 12 months, descriptive statistics are straight-forward with SAS Proc Corr.

/* Generate table of baseline (BL), 6 month (M6), 12 month (M12) means, variances, correlation matrix */

```
Proc Corr Data=Across;
Var BL_HbA1c M6_HbA1c M12_HbA1c;
Run;
```

Table 2.1: Proc Corr Summary Statistics

Variable	N	Mean	Std Dev	Min	Max
BL_HbA1c	211	10.97	1.64	8.2	14
M6_HbA1c	192	9.65	1.94	5.4	14
M12_HbA1c	176	9.97	2.10	5.9	14

When looking at the correlation matrix, I am interested in whether the correlations appear to be constant over time or attenuating over time. Constant correlation over time indicates a compound symmetry correlation structure (Table 2.2) and attenuating correlation suggests an autoregressive correlation structure (Table 2.3). It's important to always view the data first, before viewing statistics on which correlation structure best fits the data.

Let ρ = Pearson correlation coefficient between outcomes at two time points.

Table 2.2: Correlation Matrix with Compound Symmetry with Three Time Points

	Y(time = 1)	Y(time = 2)	Y(time = 3)
Y(time = 1)	1	ρ	ρ
Y(time = 2)	ρ	1	ρ
Y(time = 3)	ρ	ρ	1

Table 2.3: Correlation Matrix with Auto-Regressive AR(1) Structure

	Y(time = 1)	Y(time = 2)	Y(time = 3)
Y(time = 1)	1	ρ	ρ^2
Y(time = 2)	ρ	1	ρ
Y(time = 3)	ρ^2	ρ	1

Between every combination of two variables, SAS will display the correlation information in three lines. The first line is the correlation coefficient. The second line is the p-value and the third line is the sample size for which data are present in both comparison variables. In Table 2.4, the correlation coefficients between (baseline and month 6) and (baseline and month 12) appear to be declining, .257 and .184, suggesting compound symmetry. However, the correlation coefficient between months 6 and 12 is .283. So, the empirical correlation structure, the one based on viewing the data, appears to be autoregressive (decaying correlation).

Table 2.4: Correlation Matrix from Proc Corr

Pearson Correlation Coefficients			
Prob > |r| under H0: Rho=0			
Number of Observations			
	BL_HbA1c	**M6_HbA1c**	**M12_HbA1c**
BL_HbA1c	1	0.257	0.184
		0.0003	0.0001
	211	192	176
M6_HbA1c	0.257	1	0.283
	0.0003		<.0001
	192	192	166
M12_HbA1c	0.184	0.283	1
	0.0001	<.0001	
	176	166	176

Tip. Finding the number of measurements per person or per subject. When working with longitudinal data, it's also important to know the number of measurements per person or per subject.

Option 1. Use a dataset with baseline, 6 month, 12 month values as column variables. The N function counts the number of non-missing values.
Tip: This is the easiest method if the data has discrete time points.

```
NumHbA1c=N(BL_HbA1c, M6_HbA1c, M12_HbA1c);
```

Option 2. Use Proc Freq with the original long dataset.
Tip: Use this technique if the number of measurements per person is unknown.

```
/* NumHbA1c = number of HbA1c measurements per person */
proc freq data=LabData;
Tables ID/out=HbA1cCount;
run;

Data HbA1cCount;
Set HbA1cCount(Keep=ID Count);
Rename Count=NumHbA1c;
Run;
```

```
/* Add count to original dataset */
Data LabData;
Merge LabData HbA1cCount;
by ID;
Run;
```

Next, let's focus on how to find summary statistics and the correlation matrix when the data originally looks like the table below with varying numbers of observations per subject ID. (HbA1c = blood sugar measurement and is the outcome for this example.)

ID	HbA1cDate	HbA1c
1	2/1/2016	14
1	5/5/2016	11
1	11/7/2016	12
2	2/15/2016	13
2	8/17/2016	12
2	2/1/2017	11
2	8/5/2017	12

```
/* Add a timepoint counter to the dataset by each HbA1c measurement by ID */
Proc Sort data=LabData;
By ID HbA1cDate;
Run;

Data LabData;
Set LabData;
by ID;
TimePoint+1; /* Increment counter for each HbA1c measurement */

/* Reset counter for new ID */
If (First.ID) then TimePoint=1;
Run;
```

The dataset with the counter will look like this:

ID	TimePoint	HbA1cDate	HbA1c
1	1	2/1/2016	14
1	2	5/5/2016	11
1	3	11/7/2016	12
2	1	2/15/2016	13
2	2	8/17/2016	12
2	3	2/1/2017	11
2	4	8/5/2017	12

```
/* Then, use Proc Transpose to transpose the dataset. */
proc transpose data=LabData out=LabData_XP prefix=HbA1c;
var HbA1c;
id TimePoint;
by ID;
run;

/* View Raw Correlation Structure */
proc corr data=AcrossTime_XP;
var HbA1c1-HbA1c10;
run;
```

After the transpose, displaying the summary statistics and correlation is the same with data with discrete data.

```
proc corr data=AcrossTime_XP;
var HbA1c1-HbA1c10;
run;
```

Last, be sure to check the correlation between the outcome, Y or HbA1c, with Time, Time^2, and Time^3.

```
proc corr data=LabData;
var HbA1c;
with Time Time2 Time3;
run;
```

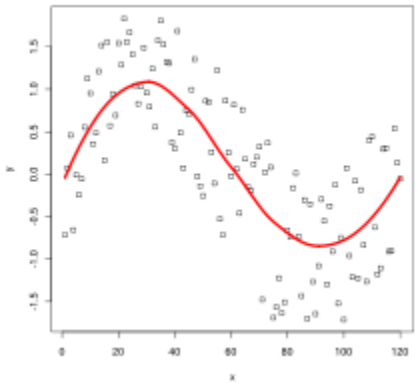
3)_Exploratory Graphics.

For exploratory graphics, I recommend a two-part approach. First, generate smoothed polynomial plots (LOESS plots) for all of the data pooled together. Please see section below on "LOESS Smoothing" for an in-depth description. Second, take a random sample from each treatment group and plot the trajectories of the outcome over time. I usually take a sample of 5-to-10 from each group and call this the spaghetti plot. The purpose of this plot is to help visualize what is happening in each treatment group over time.

LOESS Smoothing. Whether time is discrete or continuous, plotting the LOESS^2 smoothing curve is a helpful starting point. LOESS stands for Locally Weighted Scatterplot Smoothing. The LOESS technique estimates local regression polynomials over subsets of the data, given a smoothing parameter, α. The smoothing parameter, α, is also called the bandwidth and indicates the percentage of the data that is used to fit each of the local polynomials. LOESS was developed by William Cleveland and Susan Devlin.

A common choice for α is 0.5. In the graph below, the data is a sine wave and half of the data was used to fit each portion of the curve.

Figure 3.1: LOESS Curve Example with α = .5



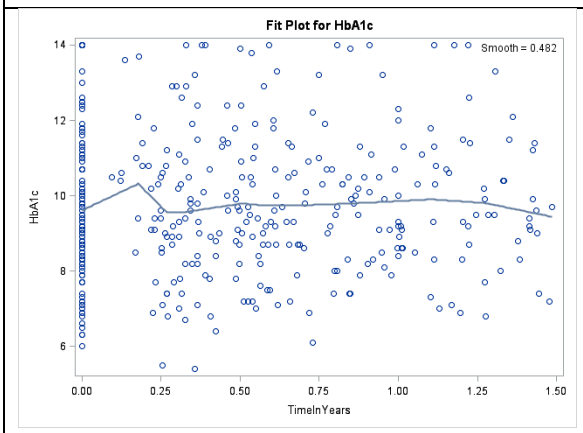
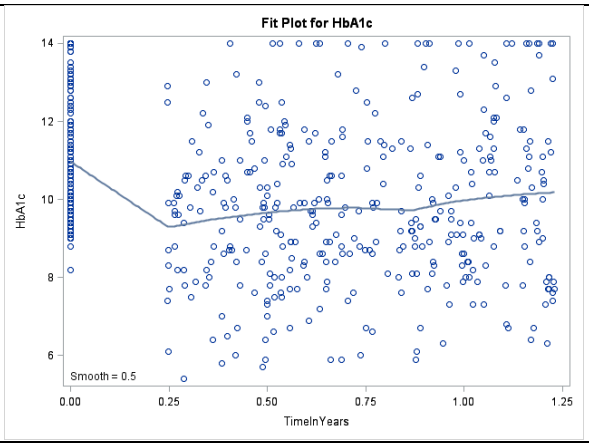
Although SAS contains algorithms to compute the optimal value of α, I like to begin by using Proc LOESS to find the optimal α.

This SAS code computes the optimum α for the treatment and control groups, indicated by treat=1 and treat=0. The degree=2 option means that the curve will be a second degree polynomial (aka parabola). A parabola was chosen because there are three time points for the treatment group's data: baseline, 6 months, and 12 months. In contrast, time was continuous for the control group. Because the control

group received usual care at the clinic, their HbA1c data was collected over a wide range of times over one year.

```
/* First, select LOESS smoothing parameter */
ods html path="c:\temp";  ods graphics on;
proc loess data=Across_long_chwpluscontrol;
   title 'PROC LOESS, Global Optimum';
   model HbA1c = TimeInYears/  degree=2 select=AICC(range(0,0.5));
   *where treat=0; /* smooth=.482 */
   where treat=1; /* smooth=.5 */
   ods select fitplot;
run;
ods graphics off; ods html close;
```

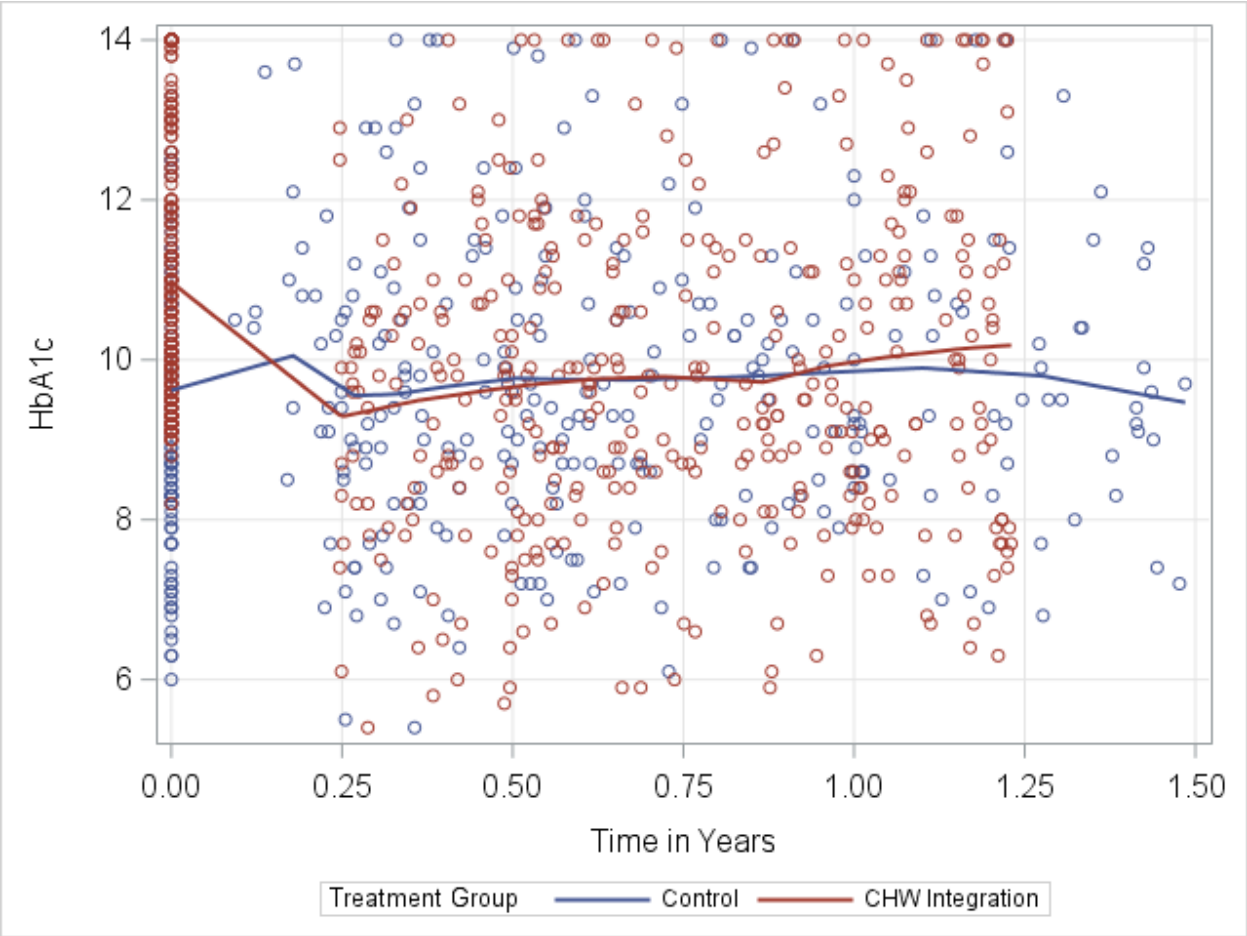
Figure 3.2: LOESS Curves and Smoothing Parameters and Fit Plots

LOESS Fit Plot: Control Group, α = 0.482	LOESS Fit Plot Treatment Group, α = 0.5
	

Next, plot the LOESS curves for the treatment and control groups on one graph by using the Group= option in Proc SGPlot.

```
/* LOESS Curves for Treatment and Control With Proc SGPLOT */
Proc SGPlot Data=Across_long_chwpluscontrol;
format Treat Trt.;
Label Treat='Treatment Group';
loess x=TimeInYears Y=HbA1c/group=treat degree=2 smooth=.5;
xaxis grid LABELATTRS=(Size=12) VALUEATTRS=(Size=12) label="Time in Years";
yaxis grid LABELATTRS=(Size=12) VALUEATTRS=(Size=12);
Run;
```

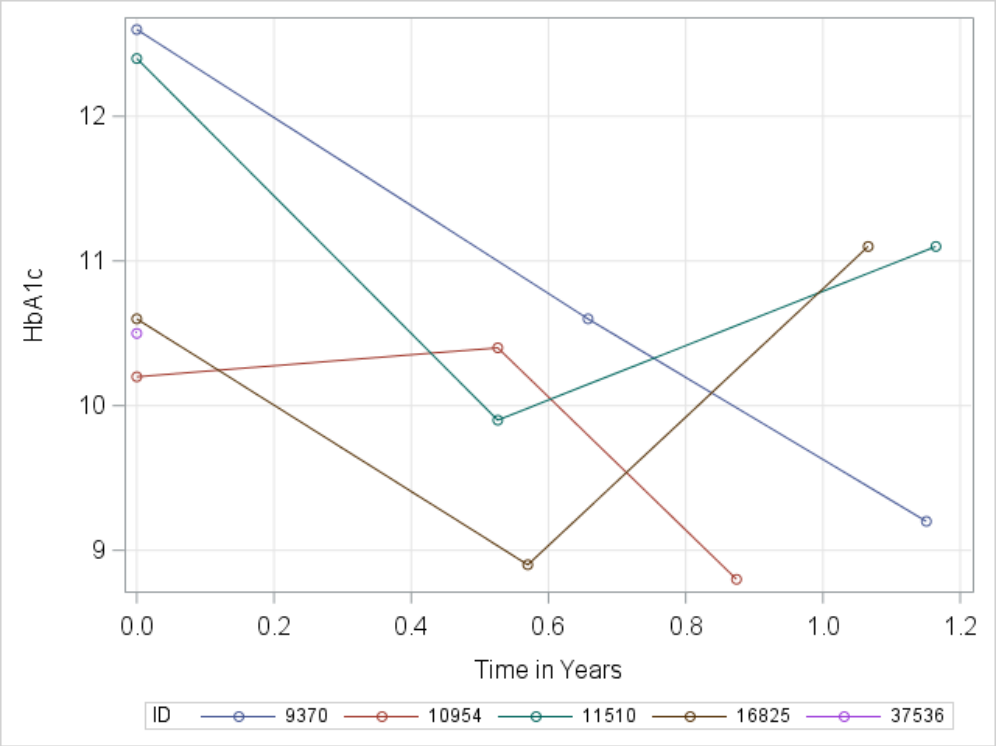
Figure 3.3: LOESS Curves and Scatter Plots for Treatment and Control



Spaghetti Plot. A spaghetti plot is a plot of a random sample of the outcomes across time for a small number of IDs.

```
/* Random selection of 5 people from baseline data */
proc surveyselect data=Baseline method=SRS N=5 Out=FiveRand seed=8152018;
run;
/* Tip: Use seed.  Otherwise, SAS will choose different 5 for procedure re-
runs */

Data FiveRandLong;
Merge FiveRand(Keep=ID IN=X) LongA1c;
by ID;
if X=0 then delete; Run;

ods html path="c:\temp"; ods graphics on;
proc sgplot data=FiveRandLong;
  series x=TimeInYears y=HbA1c / group=ID markers;
xaxis grid LABELATTRS=(Size=12) VALUEATTRS=(Size=12) label="Time in Years";
yaxis grid LABELATTRS=(Size=12) VALUEATTRS=(Size=12) Label="HbA1c";
run;
ods graphics off; ods html close;
```

Figure 3.4: Spaghetti Plot



4)_Covariance Structures in SAS® Proc Mixed.

While SAS offers a wide range of covariance structures, the most useful ones in my work are unstructured, compound symmetry, and autoregressive. Covariance structures are compared by using the AIC (Akaike Information Criteria[3]) and BIC (Bayesian Information Criteria[4]). The covariance structure that fits the data best will have the smallest AIC and BIC.

Table 3.1: Example Table with AIC and BIC from SAS Proc Mixed Example 79.2

Fit Statistics	
-2 Log Likelihood	419.5
AIC (Smaller is Better)	447.5
AICC (Smaller is Better)	452.0
BIC (Smaller is Better)	465.6

Unstructured covariance means that separate variances and coviances are estimated for each time point and for each combination of two time points. Figure 3 displays unstructured correlation three time points: baseline, 6 months, and 12 months.

Table 3.2: Unstructured Covariance Structure for Three Time Points

	Y(time = 1)	Y(time = 2)	Y(time = 3)
Y(time = 1)	σ_1^2	σ_{12}	σ_{13}
Y(time = 2)	σ_{12}	σ_2^2	σ_{23}
Y(time = 3)	σ_{13}	σ_{23}	σ_3^2

This example SAS code shows how to invoke the unstructured covariance.

```
PROC MIXED DATA = GestWt method = REML NOCLPRINT;
Class ID;
model wt = gestwk parity YearUS height triceps anthrowk / solution ddfm=kr;
repeated gestwk / subject = id type = un;
run;
```

Tips on Unstructured Covariance Structure:

- For longitudinal data with continuous time, I have encountered convergence problems with unstructured covariance on the "Repeated" statement. A structured covariance structure usually works better for continuous time.
- When time is discrete, I have seen cases where the AIC and BIC criteria indicate the unstructured covariance is the best choice.

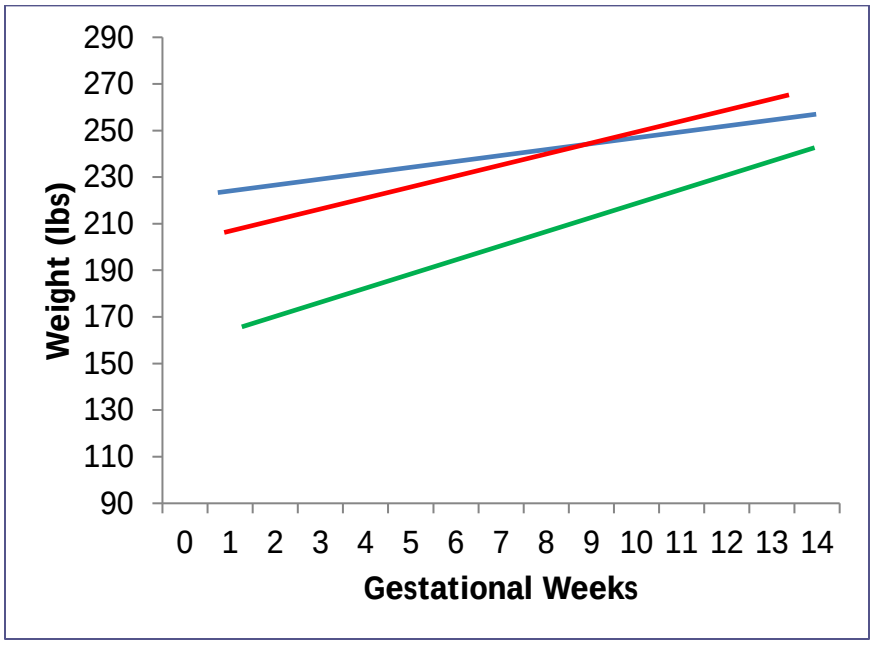
- For models with a random slope and intercept, use type=UN on the random statement. This will estimate variances for the intercept and slope, along with covariances between the intercept and slope. If "type=UN" is omitted on the random statement, the SAS default is type=VC, variance components, which assumes that the intercept and slope are independent, and estimates covariance to be zero.

In this example, gestational weights for pregnant women are being estimated.

/* The Outpred dataset estimates $Weight_{it} = \beta_0 + \beta_1 * t + b_{0i} + b_{1i} * t$ + (β's x other covariates) for each participant */

```
/* OutPred = pdat outputs person-level estimates */
PROC MIXED DATA = GestWt method = REML NOCLPRINT;
Class ID;
model wt = gestwk parity YearUS height triceps anthrowk / outpred = pdat
solution ddfm=kr;
random int gestwk / subject = id type = un g gcorr;
run;
```

Figure 3.1: Graphical View of Random Effects Model – Different Gestational Weight Trajectories for Each Woman



Tips on Compound Symmetry Covariance Structure. Equal Correlation Between Time Points

	Y(time = 1)	Y(time = 2)	Y(time = 3)
Y(time = 1)	σ^2	$\rho\sigma^2$	$\rho\sigma^2$
Y(time = 2)	$\rho\sigma^2$	σ^2	$\rho\sigma^2$
Y(time = 3)	$\rho\sigma^2$	$\rho\sigma^2$	σ^2

The above table is equivalent to this table below, where $\rho = \sigma_b^2/(\sigma_e^2 + \sigma_b^2)$.

	Y(time = 1)	Y(time = 2)	Y(time = 3)
Y(time = 1)	$\sigma_e^2 + \sigma_b^2$	σ_b^2	σ_b^2
Y(time = 2)	σ_b^2	$\sigma_e^2 + \sigma_b^2$	σ_b^2
Y(time = 3)	σ_b^2	σ_b^2	$\sigma_e^2 + \sigma_b^2$

- **First, a compound symmetry covariance structure is equivalent to the variance structure for a random intercept model.**
- Compound symmetry equivalent to random intercept model, regardless of whether time is discrete or continuous.
- Both compound symmetry and a random intercept model with produce identical AIC & BIC.

A random intercept model is defined by the equation, $Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_0 + \varepsilon_{ij}$.

- Y_{ij} = outcome for ith participant at the jth time.
- β_0 = Fixed population intercept.
- β_1 = Fixed population slope.
- t_{ij} = time for ith participant, jth measure of time.
- b_0 = Random intercept (each subject starts from different point, has their own intercept). $b_0 \sim N(0, \sigma_b^2)$.
- ε_{ij} = Random error. $\varepsilon_{ij} \sim N(0, \sigma_e^2)$.

- $Var(Y_{ij}) = Var(b_0 + \varepsilon_{ij})$. Recall from model assumptions that $Cov(b_0, \varepsilon_{ij}) = 0$. So, the diagonal elements will be $Var(b_0) + Var(\varepsilon_{ij}) = \sigma_e^2 + \sigma_b^2$.
- When $i \neq j$ or $k \neq l$, $Cov(Y_{ij}, Y_{kl}) = Cov(b_0 + \varepsilon_{ij}, b_0 + \varepsilon_{kl}) = Var(b_0) = \sigma_b^2$.
- Type=CS on the SAS repeated statement defines correlation between repeated measures on the same person with the R matrix. "Random int / Subject=ID" defines correlation between repeated measures on the same person with the G matrix and makes person-level estimation possible.

- A fixed effects model with compound symmetry covariance is invoked in SAS Proc Mixed with
- Model Y = T; Repeated /type=CS subject=ID; This model will generate only population-level estimates.

- A random intercept must be created with the Random statement.
- Model Y = T; Random int/ subject=ID;

:

- Final tip on compound symmetry. SAS has an option for compound symmetry with unequal variances between time points or repeated measures, indicated by type=CSH (Compound Symmetry Heterogenous).

	Y(time = 1)	**Y(time = 2)**	**Y(time = 3)**
Y(time = 1)	σ_1^2	$\rho\sigma_1\sigma_2$	$\rho\sigma_1\sigma_3$
Y(time = 2)	$\rho\sigma_1\sigma_2$	σ_2^2	$\rho\sigma_2\sigma_3$
Y(time = 3)	$\rho\sigma_1\sigma_3$	$\rho\sigma_2\sigma_3$	σ_3^2

Tips on Auto-regressive, Toeplitz, Spatial Covariance.

- Use `Type=AR(1)` for discrete times.
- AR(1) available when variance unequal between time points, known as auto-regressive with heterogenous variance, syntax `Type=AR(1).`
- Toeplitz covariance is a special case of auto-regressive covariance. For homogenous Toeplitz covariance, the variance on the diagonal is constant, σ^2, while the covariances for the differences between two times, j and k, are equal. $Cov(Y_{ij}, Y_{ik}) = \sigma_{|j-k|}$. For example, $Cov(Y_{i1}, Y_{i4}) = Cov(Y_{i2}, Y_{i5}) = \sigma_3$. Syntax `Type=TOEP` homogenous, `Type=TOEPH` heterogenous.
- Use spatial covariance structure for continuous times. Syntax for spatial `type=SP(POW)(Time Variable)`. Example `type=SP(POW)(TimeInYears).` If Y_{ij} and Y_{ik} differ in time by t, $Cov(Y_{ij}, Y_{ik}) = \sigma^2\rho^t$.

Autoregressive Heterogenous Covariance, ARH(1)

	Y(time = 1)	**Y(time = 2)**	**Y(time = 3)**
Y(time = 1)	σ_1^2	$\rho\sigma_1\sigma_2$	$\rho^2\sigma_1\sigma_3$
Y(time = 2)	$\rho\sigma_1\sigma_2$	σ_2^2	$\rho\sigma_2\sigma_3$
Y(time = 3)	$\rho^2\sigma_1\sigma_3$	$\rho\sigma_2\sigma_3$	σ_3^2

Toeplitz Covariance, TOEP

	Y(time = 1)	**Y(time = 2)**	**Y(time = 3)**
Y(time = 1)	σ^2	σ_1	σ_2
Y(time = 2)	σ_1	σ^2	σ_1
Y(time = 3)	σ_2	σ_1	σ^2

There are many other covariance structures, both for discrete and continuous times. More information is available in the online SAS documentation at http://documentation.sas.com/?docsetId=statug&docsetTarget=statug_mixed_syntax14.htm&docsetVersion=14.3&locale=en#statug.mixed.repeatedstmt_type.

5)_Linear Mixed Models with Discrete Time Points.

Equation for LMM with 3 discrete time points and 2 treatment groups.
$Y_{ij} = \beta_0 + \beta_1 G_i + \beta_2 t_1 + \beta_3 t_2 + \beta_4 G_i t_1 + \beta_5 G_i t_2 + \varepsilon_{ij}$.

G_i = treatment group (1 = intervention; 0 = control) for the ith participant.
t_1 = first follow-up time, often 6 months (1 = 2nd follow-up time, 0 = other time).
t_2 = second follow-up time, often 12 months (2 = 2nd follow-up time, 0 = other time).
ε_{ij} = Random error. $\varepsilon_{ij} \sim N(0, \sigma_e^2)$.

implementation in SAS.
First, merge the datasets from the various time points. I always create difference scores (aka delta) and recommend this. If there is no missing data at baseline, creating the difference score is an excellent option for reducing skewness. For a proof that differencing reduces skewness, please see appendix 1.

```
Data Across;
Merge Demographics
BaseLab.BL_Lab(keep=ID BL_HbA1cDate BL_HbA1c)      /* Baseline */
M6Lab.M6_Lab(keep=ID M6_HbA1cDate M6_HbA1c)       /* Month 6 */
M12Lab.M12_Lab(keep=ID M12_HbA1cDate M12_HbA1c);   /* Month 12 */
by ID;

/* Compute Deltas (Change scores in HbA1c */
M6BL_HbA1c = M6_HbA1c - BL_HbA1c;
M12BL_HbA1c = M12_HbA1c - BL_HbA1c;
Run;
```

Next, convert the “across” dataset to a long dataset for SAS Proc Mixed with discrete time points. TimePointN is used, because SAS sets the largest value to the reference by default. If time points 1, 2, 3, etc. are -1, -2, -3, etc., SAS will automatically set 0 to the default.

```
/* Baseline */
TimePoint=0;   TimePointN=0;
HbA1c=BL_HbA1c;
Delta_HbA1c=0;
Output;

/* 6 Months */
TimePoint=1; TimePointN=-1;
HbA1c=M6_HbA1c;
Delta_HbA1c=M6BL_HbA1c;
Output;

/* 12 Months */
TimePoint=2; TimePointN=-2;
HbA1c=M12_HbA1c;
Delta_HbA1c=M12BL_HbA1c;
Output;
Run;

ods html; ods graphics on;
Proc Mixed Data=HbA1cLong_Discrete Method=REML NOCLPRINT
plots(only)=(StudentPanel(conditional box));
Class ID TreatN TimePointN;
```

```
Model HbA1c= TreatN TimePointN TimePointN*TreatN/ Solution
Influence(effect=ID Est) ddfm=KR;
Repeated / type=AR(1) Subject=ID R RCorr;
Where HbA1c NE .;
Estimate 'Control BL' Int 1 TreatN 0 1 TimePointN 0 0 1 TimePointN*TreatN 0 0
0 0 0 1;
Estimate 'CHWInt BL' Int 1 TreatN 1 0  TimePointN 0 0 1 TimePointN*TreatN 0 0
1 0 0 0;
Estimate 'Control M6' Int 1 TreatN 0 1 TimePointN 0 1 0 TimePointN*TreatN 0 0
0 0 1 0;
Estimate 'CHWInt M6' Int 1 TreatN 1 0  TimePointN 0 1 0 TimePointN*TreatN 0 1
0 0 0 0;
Estimate 'Control M6-BL'  TimePointN 0 1 -1  TimePointN*TreatN 0 0 0 0 1 -1;
Estimate 'CHWInt M6-BL'  TimePointN 0 1 -1  TimePointN*TreatN 0 1 -1 0 0 0;
Estimate 'Int Eff M6'  TimePointN*TreatN 0 1 -1 0 -1 1;
Estimate 'Control M12' Int 1 TreatN 0 1 TimePointN 1 0 0 TimePointN*TreatN 0
0 0 1 0 0;
Estimate 'CHWInt M12' Int 1 TreatN 1 0  TimePointN 1 0 0 TimePointN*TreatN 1
0 0 0 0 0;
Estimate 'Control M12-BL'  TimePointN 1 0 -1  TimePointN*TreatN 0 0 0 1 0 -1;
Estimate 'CHWInt M12-BL' TimePointN 1 0 -1  TimePointN*TreatN 1 0 -1 0 0 0;
Estimate 'Int Eff M12'  TimePointN*TreatN 1 0 -1 -1 0 1;
Store MixARDiscrete;  Run;
ods graphics off; ods html close;

/* Adjust Multiple Comparisons – PLM generates original & adjusted confidence
intervals.  Use seed to ensure same results on procedure re-runs. */
ods html; ods graphics on;
PROC PLM Restore=MixARDiscrete;
Estimate 'Control BL' Int 1 TreatN 0 1 TimePointN 0 0 1 TimePointN*TreatN 0 0
0 0 0 1 /adjust=simulate(NSAMP=10000 SEED=8132018);
Estimate 'CHWInt BL' Int 1 TreatN 1 0  TimePointN 0 0 1 TimePointN*TreatN 0 0
1 0 0 0 /adjust=simulate(NSAMP=10000 SEED=8132018);
… Same estimate statements as in Proc Mixed ***;
ODS Output Estimates=PLMEstARDiscrete;  Run;

Proc Print Data=PLMEstARDiscrete; Run;
ods graphics off; ods html close;
```

6)_Linear Mixed Models When Time Is Continuous.

Equation for LMM with continuous time and 2 treatment groups.

$Y_{ij} = \beta_0 + \beta_1 G_i + \beta_2 t_{ij} + \beta_3 t_{ij}^2 + \beta_4 G_i t_{ij} + \beta_5 G_i t_{ij}^2 + \varepsilon_{ij}$.

G_i = treatment group (1 = intervention; 0 = control) for the ith participant.
t = time for ith participant at jth repetition.
ε_{ij} = Random error. $\varepsilon_{ij} \sim N(0, \sigma_e^2)$.

First, merge the datasets from the various time points.

```
Data Across;
Merge Demographics
BaseLab.BL_Lab(keep=ID BL_HbA1cDate BL_HbA1c)      /* Baseline */
M6Lab.M6_Lab(keep=ID M6_HbA1cDate M6_HbA1c)       /* Month 6 */
M12Lab.M12_Lab(keep=ID M12_HbA1cDate M12_HbA1c);   /* Month 12 */
by ID;
```

```
/* Compute Time Differences */
BL_TimeInDays = 0;
M6_TimeInDays = datdif(BL_HbA1cDate, M6_HbA1cDate, 'act/act');
M12_TimeInDays = datdif(BL_HbA1cDate, M12_HbA1cDate, 'act/act');

M6_TimeInMonths = M6_TimeInDays/30;
M12_TimeInMonths = M12_TimeInDays/30;

M6_TimeInYears = M6_TimeInDays/365;
M12_TimeInYears = M12_TimeInDays/365;

/* Compute Deltas (Change scores in HbA1c */
M6BL_HbA1c = M6_HbA1c - BL_HbA1c;
M12BL_HbA1c = M12_HbA1c - BL_HbA1c;
Run;
```

Next, convert the “across” dataset to a long dataset for SAS Proc Mixed.

```
/* Baseline */
TimeInYears=BL_TimeInYears;
TimeInDays=BL_TimeInDays;
TimeInMonths=BL_TimeInMonths;
HbA1c=BL_HbA1c;
Delta_HbA1c=0;
Output;

/* 6 Months */
TimeInYears=M6_TimeInYears;
TimeInDays=M6_TimeInDays;
TimeInMonths=M6_TimeInMonths;
HbA1c=M6_HbA1c;
Delta_HbA1c=M6BL_HbA1c;
Output;

/* 12 Months */
TimeInYears=M12_TimeInYears;
TimeInDays=M12_TimeInDays;
TimeInMonths=M12_TimeInMonths;
HbA1c=M12_HbA1c;
Delta_HbA1c=M12BL_HbA1c;
Output;
Run;
```

/* Proc Mixed Code to Estimate HbA1c, 2nd Degree Polynomial, Compare 2 Treatment Groups */

```
ods html; ods graphics on;
Proc Mixed Data=LongA1c.Hba1c_long_chwpluscontrol Method=REML NOCLPRINT
plots(only)=(StudentPanel(conditional box));
Class ID TreatN;
Model HbA1c= TreatN TimeInYears TimeInYears2 TreatN*TimeInYears
TreatN*TimeInYears2 AgeGE55
  / Solution Influence(effect=ID Est) ddfm=KR;
Repeated / type=SP(Pow)(TimeInYears) Subject=ID R RCorr;
Where HbA1c NE .;
```

```
Estimate 'Control BL' Int 1 TreatN 0 1 TimeInYears 0 TimeInYears2 0 AgeGE55
.426/cl;
Estimate 'CHWInt BL' Int 1 TreatN 1 0 TimeInYears 0 TimeInYears2 0 AgeGE55
.426/cl;
Estimate 'Control M6' Int 1 TreatN 0 1 TimeInYears .5 TimeInYears2 .25
TreatN*TimeInYears 0 .5 TreatN*TimeInYears2 0 0.25 AgeGE55 .426/cl;
Estimate 'CHWInt M6' Int 1 TreatN 1 0 TimeInYears .5  TimeInYears2 .25
TreatN*TimeInYears .5 0 TreatN*TimeInYears2 0.25 0 AgeGE55 .4268/cl;
Estimate 'Control M6-BL' TimeInYears .5  TimeInYears2 .25  TreatN*TimeInYears
0 .5 TreatN*TimeInYears2 0 .25/cl;
Estimate 'CHWInt M6-BL'  TimeInYears .5  TimeInYears2 .25  TreatN*TimeInYears
.5 0 TreatN*TimeInYears2 .25 0/cl;
Estimate 'Int Eff M6' TreatN*TimeInYears .5 -.5 TreatN*TimeInYears2 .25 -
.25/cl;
Estimate 'Control M12' Int 1 TreatN 0 1 TimeInYears 1 TimeInYears2 1
TreatN*TimeInYears 0 1 TreatN*TimeInYears2 0 1 AgeGE55 .426/cl;
Estimate 'CHWInt M12' Int 1 TreatN 1 0 TimeInYears 1  TimeInYears2 1
TreatN*TimeInYears 1 0 TreatN*TimeInYears2 1 0 AgeGE55 .426/cl;
Estimate 'Control M12-BL' TimeInYears 1  TimeInYears2 1  TreatN*TimeInYears 0
1 TreatN*TimeInYears2 0 1/cl;
Estimate 'CHWInt M12-BL'  TimeInYears 1  TimeInYears2 1  TreatN*TimeInYears 1
0 TreatN*TimeInYears2 1 0/cl;
Estimate 'Int Eff M12' TreatN*TimeInYears 1 -1 TreatN*TimeInYears2 1 -1/cl;
ODS Output Estimates=HbA1cAdjEstimates;
Store HbA1cTimeCont;
Run;
ods graphics off; ods html close;

/* Adjust Multiple Comparisons – PLM generates original & adjusted confidence
intervals */
ods html; ods graphics on;
PROC PLM Restore=HbA1cTimeCont;
Estimate 'Int Eff M12' TreatN*TimeInYears 1 -1 TreatN*TimeInYears2 1 -1/
/adjust=simulate(NSAMP=10000 SEED=8132018)cl;
… Same estimate statements as in Proc Mixed ***;
ODS Output Estimates=PLMEst;
Run;

Proc Print Data=PLMEst; Run;
ods graphics off; ods html close;
```

Tip – Inestimable Error on Estimate Statements. When I worked with a model that had HbA1c data over 5 years, I originally coded time in months and the model had a 3rd degree polynomial with SP(POW)(TimeInMonths) covariance structure. Although the model converged, I was getting "Inestimable" errors from the estimate statements. I spent over an hour examining the estimate statements, but could not find any problems. So, I called SAS tech support and they advised me that the large (Time in Months)^3 values combined with small coefficient values can cause inestimable errors. So, I revised the model to use TimeInYears. The model with TimeInYears produced the same AIC and BIC values, but the estimate statements no longer produced the inestimable errors.

Table 6.1: HbA1c Results by Group and Time Point, Mean (95%CI)

Group	Baseline	6 Months	12 Months
Control	9.64 (9.29, 9.99)	9.82 (9.48, 10.17)	9.86 (9.49, 10.23)
CHW Intervention	10.95 (10.69, 11.20)	9.63 (9.37, 9.90)	9.82 (9.55, 10.09)

Table 6.2: HbA1c Change Scores by Group and Time Point, Mean (95% CI)

Group	6 Months	12 Months
Control	0.18 (-0.15, 0.51)	0.22 (-0.22, 0.66)
CHW Intervention	**-1.31 (-1.57, -1.06)**	**-1.12 (-1.44, -0.80)**
Intervention Effect	**-1.49 (-1.91, -1.07)**	**-1.35 (-1.89, -0.80)**

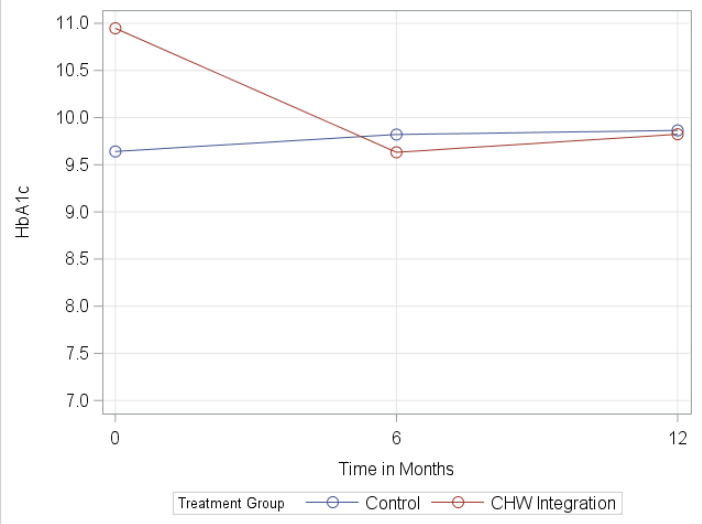
7)_Graphical Presentation of Results.

Tip. If the data is in the format: Y_Group1, Y_Group2, Time, the graph options are easier to control than if the data is in the format: Y, Group, Time.

Example 1: Data format: Y, Group, Time. Group= option triggers SAS defaults, which are modifiable by editing the graph template – a bit complicated.

```
proc sgplot data=HbA1cAdjTime;
Series x=TimeInMonths y=HbA1c/markers  markerattrs=(size=12) Group=Treat;
xaxis grid values=(0 to 12 by 6) LABELATTRS=(Size=12) VALUEATTRS=(Size=12);
yaxis  grid  label='HbA1c'  values=(7  to  11  by  .5)  LABELATTRS=(Size=12)
VALUEATTRS=(Size=12);
keylegend /  VALUEATTRS=(Size=12);
run;
```

Figure 7.1: Outcome Trajectory Using SAS Default Group= Settings

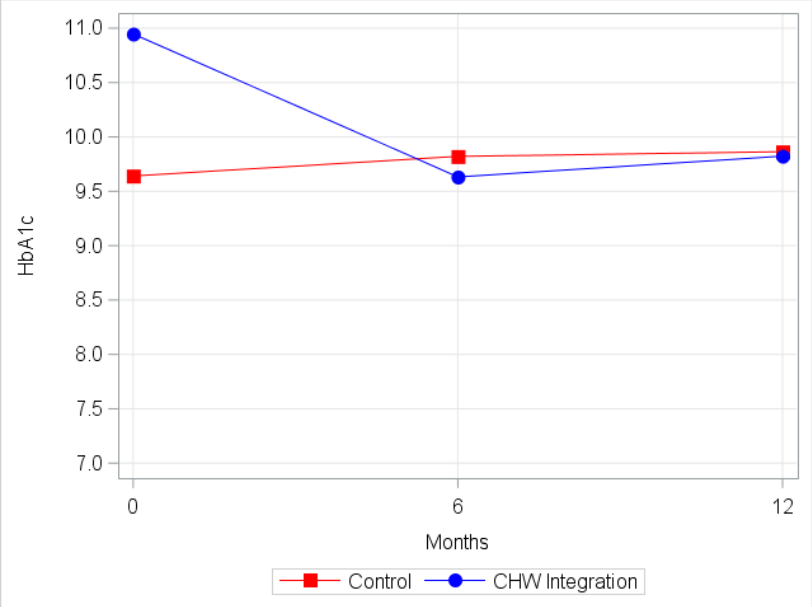


Example 2: Data Format: Y_Group1, Y_Group2, Time. Easier to customize graph with options on the Series statement.

```
proc sgplot data=HbA1cAdjTimeBoth;

/* Must have markers option on Series statement.  Otherwise, markattrs will
be ignored. */
Series x=TimeInMonths y=HbA1cCtrl/markers lineattrs=(color=red)
markerattrs=(size=12 symbol=SquareFilled color=red);
Series x=TimeInMonths y=HbA1cTrt/markers lineattrs=(color=blue)
markerattrs=(size=12 symbol=CircleFilled color=blue);
```

```
xaxis grid label='Months' values=(0 to 12 by 6) LABELATTRS=(Size=12)
VALUEATTRS=(Size=12);
yaxis grid label='HbA1c' values=(7 to 11 by .5) LABELATTRS=(Size=12)
VALUEATTRS=(Size=12);
keylegend /  VALUEATTRS=(Size=12);
run;
```

Figure 7.2: Outcome Trajectory Using Custom Colors and Attributes



Another option is to create a bar chart for changes at specific times relative to baseline with 95% confidence intervals. $\Delta Y_t = Y_t - Y_0$; change score = Y at time t; Y_0 at time 0 (baseline).

Tip. When displaying a graph of change scores, using Vbar or Xbar in Proc SGPlot tells SAS to calculate confidence intervals from the data. Using the Scatter statement allows use of a dataset with pre-calculated confidence limits from a previous SAS procedure, such as Proc Mixed. Suppose the output estimates from Proc Mixed have the format below, where lower and upper represent the 95% confidence interval for ΔY.

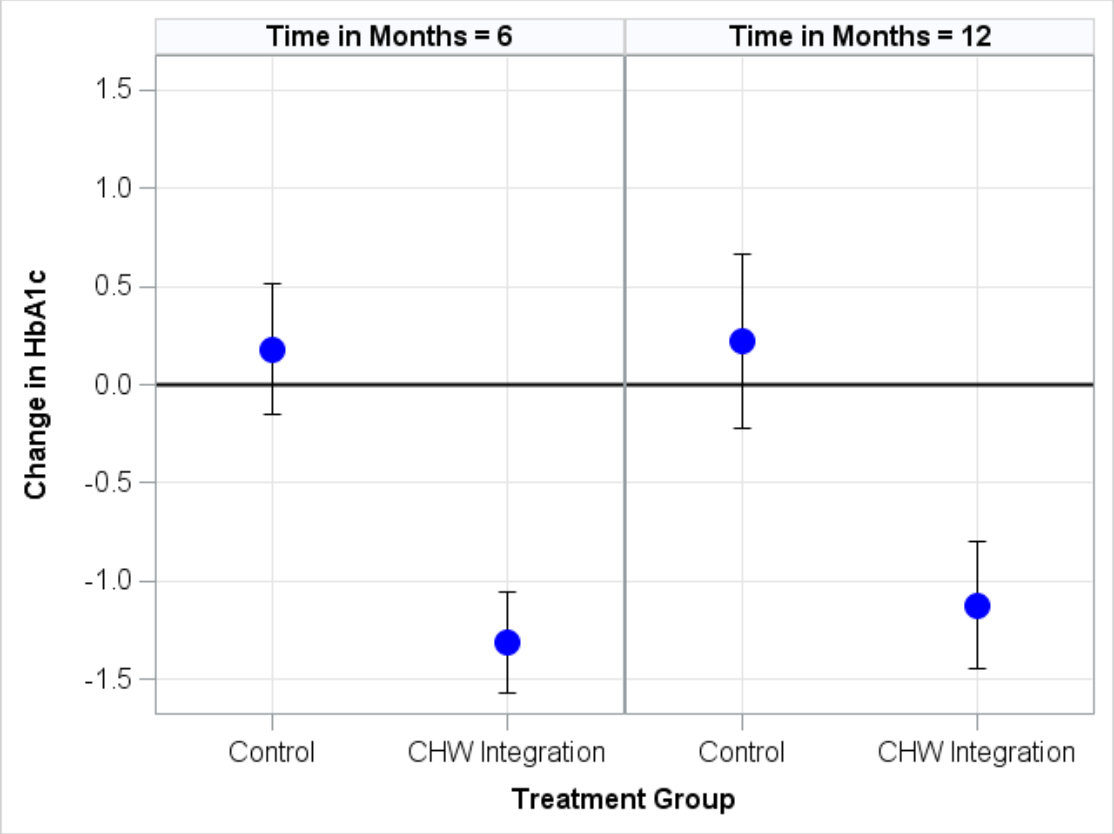
Group	TimeInMonths	DeltaY	Lower	Upper
Control	6	0.18	-0.15	0.51
CHW Integration	6	-1.31	-1.57	-1.05
Control	12	0.22	-0.22	0.67
CHW Integration	12	-1.12	-1.44	-0.80

This SAS code will produce a nice graph of ΔY at 6 and 12 month follow-up.

```
proc sgpanel data=DeltaHbA1cAdjTime;

panelby TimeInMonths / rows=1 columns=2 HEADERATTRS=(Color=Black Size=12
Weight=Bold);

scatter x=Treat y=DeltaHbA1c /  ERRORBARATTRS=(color=black) yerrorlower=Lower
yerrorupper=Upper markerattrs=(symbol=circlefilled color=blue size=20);

colaxis grid LABELATTRS=(Size=12 weight=bold) VALUEATTRS=(Size=12) values=(0
1) type=discrete offsetmin=.25 offsetmax=.25;
```

```
rowaxis grid LABELATTRS=(Size=12 weight=bold) VALUEATTRS=(Size=12) values=(-
1.5 to 1.5 by .5);
/* Reference line at Y=0 */
refline 0/ axis=y lineattrs=(Color=Black Thickness=2);
run;
```

Figure 7.3: Change Scores by Treatment Group and Time Point



8)_Conclusions.

- Begin with LOESS and spaghetti plots, whether time points are discrete or continuous.
- Explore empirical correlation matrices, means, and standard deviation with Procs Corr and Means.
- SAS syntax and correlation structures differ between linear mixed models with discrete and continuous times.
- Find the best fitting covariance structure with AIC and BIC.
- “Inestimable” errors on estimate statement can sometimes be fixed by changing the unit of time.
- Use Proc PLM to adjust for multiple comparisons with Monte Carlo simulation.
- Use the seed option for any procedures that draw a random sample, such as SurveySelect or PLM.
- When creating graphics, best to choose graphics where options are easily changeable without modifying the graph template. I.E., easiest to customize the graph with axis and marker options within the SG procedures.

9)_Acknowledgements.

- University of Michigan Schools of Social Work and Public Health.
- Community Health and Social Services (CHASS) of Detroit, MI.
- Funding: National Institute for Health Care Reform (nihcr.org).

REFERENCES

1. West BT, Welch KB, Galecki AT. *Linear mixed models: A practical guide using statistical software.* 2nd ed. New York, NY: Chapman & Hall/CRC; 2014.

2. Cleveland WS, Devlin SJ. Locally-weighted regression: An approach to regression analysis by local fitting. *JASA*. 1988;83(403):596-610.

3. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974;AC-19:716-723.

4. Schwarz G. Estimating the dimension of a model. *Annals of Statistics*. 1978;6:461-464.

CONTACT INFORMATION

Your comments and questions are welcome.

Brandy R. Sinco, Statistician and Programmer/Analyst
University of Michigan School of Social Work
1080 S. University St.
Box 183
Ann Arbor, MI 48109-1106

Phone: 734-763-7784

E-Mail: brsinco@umich.edu