Paper 5800-2016

A Short Introduction to Longitudinal and Repeated Measures Data Analyses

Leanne Goldstein, City of Hope

ABSTRACT

Longitudinal and repeated Measures data are seen in nearly all fields of analysis. Examples of this data include weekly lab test results of patients or test scores by children from the same class. Statistics students and analysts alike may be overwhelmed when it comes to repeated measures or longitudinal data analyses. They may try to educate themselves by diving into text books or taking semester long or intensive weekend courses resulting in even more confusion. Some may try to ignore the repeated nature of data and take short cuts such as analyzing all data as independent observations or analyzing summary statistics such as averages or changes from first to last points and ignoring all the data in-between. This Hands-On presentation will introduce longitudinal and repeated measures analyses without heavy emphasis on theory. Students in the workshop will have the opportunity to get hands-on experience graphing longitudinal and repeated measures data. They will learn how to approach these analyses with tools like PROC MIXED and PROC GENMOD. Emphasis will be on continuous outcomes but categorical outcomes will briefly be covered.

INTRODUCTION

In covering longitudinal and repeated measures data, the primary idea is to understand what it is and why use it. We can think of starting from the framework of Statistics 101 courses where students become familiar with Analyses of Variance (ANOVA) and simple linear regression models. They are taught that among the basic assumptions necessary to perform these analyses, observations need to be independent, i.e. not correlated. But what happens when there is a violation of this assumption? What happens when subjects contribute more than one observation to a sample and we want to analyze all of the data? Some analysts may dangerously proceed without consideration of violating the independence assumption. Others may choose to summarize the repeated data points on a subject, analyzing only the mean or minimum or maximum or change. While the latter may be a valid form of analysis, the analyst will lose out on identifying interesting data patterns and relationships. Therefore, it is important to consider longitudinal or repeated measures data analysis.

Students often get overwhelmed with learning about longitudinal data analysis because it involves new unfamiliar SAS procedures, i.e. PROC MIXED, PROC GENMOD or PROC GLIMMIX. Statistical textbooks or courses may overwhelm the novice statistician with the content of many formulas and considerations. This HOW will cover introductory concepts of repeated measures analyses to get the novice statistician going.

BACKGROUND

The key concept of repeated measures or longitudinal analyses that needs to be understood is that repeated measures analyses has to do with partitioning model error. Novice statisticians should be familiar with a basic linear regression or analysis of variance formula

Y = XB + E where

Y represents the outcome or dependent variable

X represents the independent variables (covariates),

B is the regression or ANOVA coefficients (parameter estimates)

E is the error (residual) or the difference between the model-predicted value and the outcome

In a repeated measures analysis, when we have a violation of independence assumption we wish to partition the error into subject error called D and overall error called E. Therefore, the new equation for a repeated measures analysis will be

Y = XB + D + E

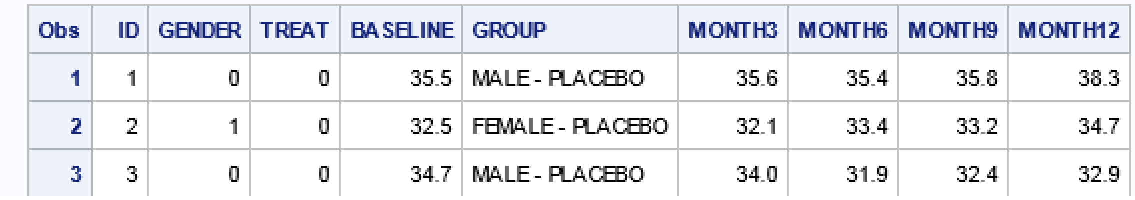
Note how the X and B were not affected. This partition makes XB the fixed effects of the model and D the random effects of the model. The main take away is that in repeated measures analyses, we still test covariates, interactions, etc. as we normally would in ordinary linear regression, we are just accounting for partitioning the error.

SETTING UP DATA FOR ANALYSES

For this workshop we are going to use a simulated data set of Body Mass Index (BMI) measurements from subjects on an imaginary weight loss study. The advantage of simulated data is that we know the answer ahead of time, but the disadvantage is that we aren't getting a real picture of how messy real data really are. For simplicity, suppose that we randomize 100 patients with 50 males (GENDER=0) and 50 females (GENDER=1) to either the weight loss drug (TREAT=1) or placebo (TREAT=0). Suppose that we track their BMI at baseline and 3, 6, 8, and 12 months on study. The GROUP variable will explain the gender and treatment combination.

```
DATA BMI;
 CALL STREAMINIT(12345);
      DO ID = 1 TO 100;
      GENDER=(MOD(ID,2)=0);
      TREAT=( ID>50);
      BASELINE = ROUND(RAND('NORMAL',35,2),.1);
      IF GENDER=1 AND TREAT=0 THEN DO;
            GROUP = 'FEMALE - PLACEBO';
            MONTH3 =  ROUND(BASELINE - .25 + RAND('NORMAL',0,1),.1);
            MONTH6 =  ROUND(MONTH3  + .25  + RAND('NORMAL',0,1),.1);
            MONTH9 =  ROUND(MONTH6  - .25  + RAND('NORMAL',0,1),.1);
            MONTH12=  ROUND(MONTH9  + .25  + RAND('NORMAL',0,1),.1);
            END;
      IF GENDER=0 AND TREAT=0 THEN DO;
            GROUP = 'MALE - PLACEBO';
            MONTH3 =  ROUND(BASELINE - 1 + RAND('NORMAL',0,1),.1);
            MONTH6 =  ROUND(MONTH3  - 1 + RAND('NORMAL',0,1),.1);
            MONTH9 =  ROUND(MONTH6  + 1 + RAND('NORMAL',0,1),.1);
            MONTH12=  ROUND(MONTH9  + 1 + RAND('NORMAL',0,1),.1);
            END;
    IF GENDER=0 AND TREAT=1 THEN DO;
            GROUP = 'MALE - TREAT';
            MONTH3 =  ROUND(BASELINE - 1.5 + RAND('NORMAL',0,1),.1);
            MONTH6 =  ROUND(MONTH3  - 1.5 + RAND('NORMAL',0,1),.1);
            MONTH9 =  ROUND(MONTH6  - 1.5 + RAND('NORMAL',0,1),.1);
            MONTH12=  ROUND(MONTH9  - 1.5 + RAND('NORMAL',0,1),.1);
            END;
      IF GENDER=1 AND TREAT=1 THEN DO;
            GROUP = 'FEMALE - TREAT';
            MONTH3 =  ROUND(BASELINE - 1 + RAND('NORMAL',0,1),.1);
            MONTH6 =  ROUND(MONTH3  - 1 + RAND('NORMAL',0,1),.1);
            MONTH9 =  ROUND(MONTH6  - 1 + RAND('NORMAL',0,1),.1);
            MONTH12=  ROUND(MONTH9  - 1 + RAND('NORMAL',0,1),.1);
            END;
      OUTPUT;
      END;
      RUN;
```

If we print out this data set, we note that the data is in what we call WIDE format. This means that there is one subject per row and the multiple time points are collected in different columns. Unless you are working with pharmaceutical data or other well structured data sets, longitudinal study data are often collected in spreadsheets in WIDE format. Output 1 shows the appearance of printing out the wide BMI data set.



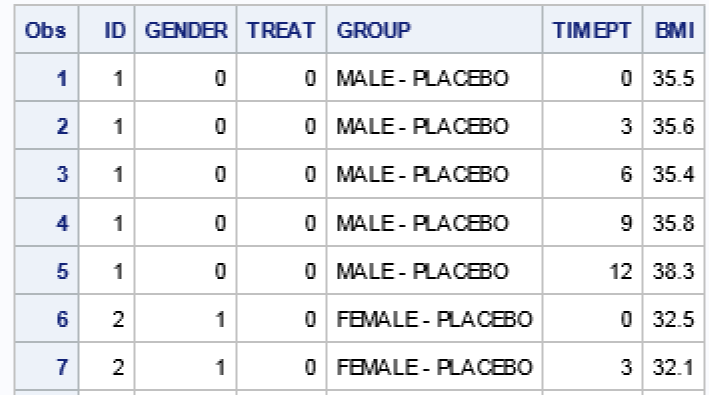
Obs	ID	GENDER	TREAT	BASELINE	GROUP	MONTH3	MONTH6	MONTH9	MONTH12
1	1	0	0	35.5	MALE - PLACEBO	35.6	35.4	35.8	38.3
2	2	1	0	32.5	FEMALE - PLACEBO	32.1	33.4	33.2	34.7
3	3	0	0	34.7	MALE - PLACEBO	34.0	31.9	32.4	32.9

Output 1. Output from simulated BMI data set in WIDE format

While it is visually pleasing to look across the columns and see the changes in BMI, for analysis purposes we need to convert this data to LONG format. This means all the BMI responses should be in one column with another column which labels each response's time point. The following code converts this data into LONG format with all responses contained in the variable BMI and all corresponding time points contained in the variable TIMEPT. BASELINE will naturally get labeled TIMEPT=0. We set TIMEPT to the appropriate time point and BMI to the appropriate variable and then OUTPUT the result. We drop the BASELINE variable and all the MONTH variables because these are no longer necessary for the analytic data set. Note that instead of listing out each MONTH variable we can tell SAS to drop all variables that start with MONTH by using the semi-colon after MONTH, i.e. MONTH:.

```
DATA BMILONG;
      SET BMI;
      TIMEPT=0; BMI=BASELINE; OUTPUT;
      TIMEPT=3; BMI=MONTH3; OUTPUT;
      TIMEPT=6; BMI=MONTH6; OUTPUT;
      TIMEPT=9; BMI=MONTH9; OUTPUT;
      TIMEPT=12; BMI=MONTH12; OUTPUT;
      DROP BASELINE MONTH:;
      RUN;
```

Output 2 shows the printed out LONG data set with all the responses in BMI and TIMEPT indicating which month each BMI was taken and ID identifying the subject. This LONG data layout has multiple rows for each subject and is not ideal for visually digesting the patterns of change in BMI, but it is the way the data need to be for repeated measures analyses.



Obs	ID	GENDER	TREAT	GROUP	TIMEPT	BMI
1	1	0	0	MALE - PLACEBO	0	35.5
2	1	0	0	MALE - PLACEBO	3	35.6
3	1	0	0	MALE - PLACEBO	6	35.4
4	1	0	0	MALE - PLACEBO	9	35.8
5	1	0	0	MALE - PLACEBO	12	38.3
6	2	1	0	FEMALE - PLACEBO	0	32.5
7	2	1	0	FEMALE - PLACEBO	3	32.1

Output 2. Output from BMI data set in LONG format

GRAPHING LONGITUDINAL DATA

Before embarking on any analyses of longitudinal data or repeated measures data, be sure that the data are properly sorted. We want the data sorted by subject and then by time. This may be redundant here but we show you this step so that you are reminded to sort when you use other data for this type of analysis.

```
        PROC SORT DATA=BMILONG; BY ID TIMEPT; RUN;
```

Now that the data are ready for analysis, let's plot it so that we can get an idea of what the BMI trends look like. The following procedure generates what is called a spaghetti plot of the data, so called because all the lines will make it look like spaghetti. We use the SERIES statement in PROC SGPLOT with TIMEPT on the horizontal x axis and BMI on the vertical y axis. We wish to have one line per person so we use GROUP=ID to do so. For simplicity we specify options LINEATTRS with black solid (pattern=1) line and mark each point with an asterisk in the MARKERATTRS option. We want to create tick marks every three months and indicate this in the XAXIS statement. To modify the graph attributes such as the axes labels and colors, please see PROC SGPLOT SAS documentation.

```
PROC SGPLOT DATA=BMILONG;
        SERIES X=TIMEPT Y=BMI / GROUP=ID LINEATTRS= (COLOR=BLACK PATTERN=1)
        MARKERS MARKERATTRS=(SYMBOL=ASTERISK);
        XAXIS VALUES=(0 TO 12 BY 3);
RUN;
```

Figure 1 shows the output from the PROC SGPLOT. This figure is really useful and informative to both the analyst and the investigator because it tells the story of the entire cohort. From the figure we see that all patients started at a BMI ranging from 30 to 40 which is by definition obese. Overall, the cohort BMI doesn't change much at month 3 as the range in BMI is still approximately 30 to 40. Thereafter, at months 6, 9, and 12, patients BMIs either go up or down. At the end of 12 months, some patients have a much lower BMI than what they started with and others have an increase in BMI (range of BMIs at 12 months is 20 to over 40). A summary plot which would just show means and ranges at each time point could provide this information too, but a Spaghetti plot is useful to somewhat see individual subject trends in BMI. The PROC SGPLOT could be run with a BY statement as well to look at patterns for each treatment or placebo group of by gender.

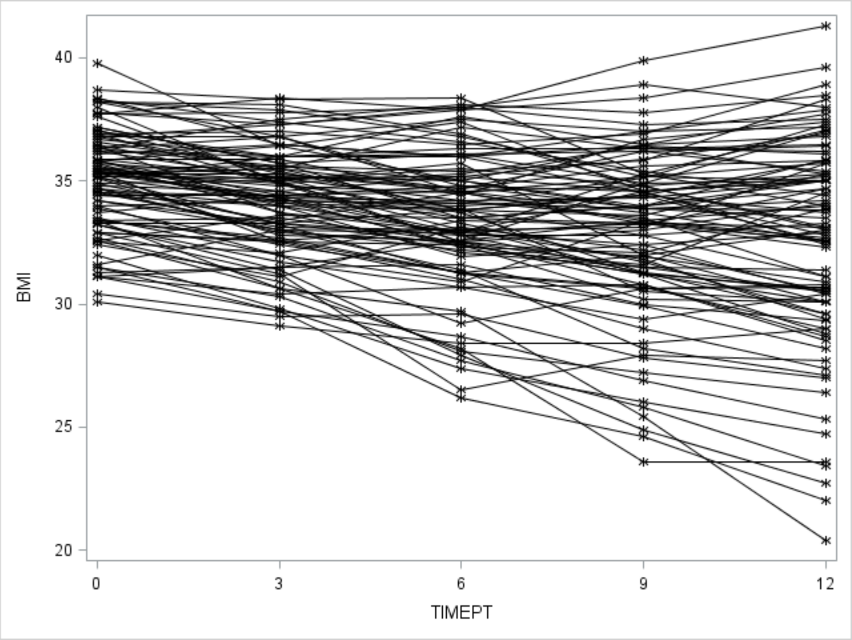


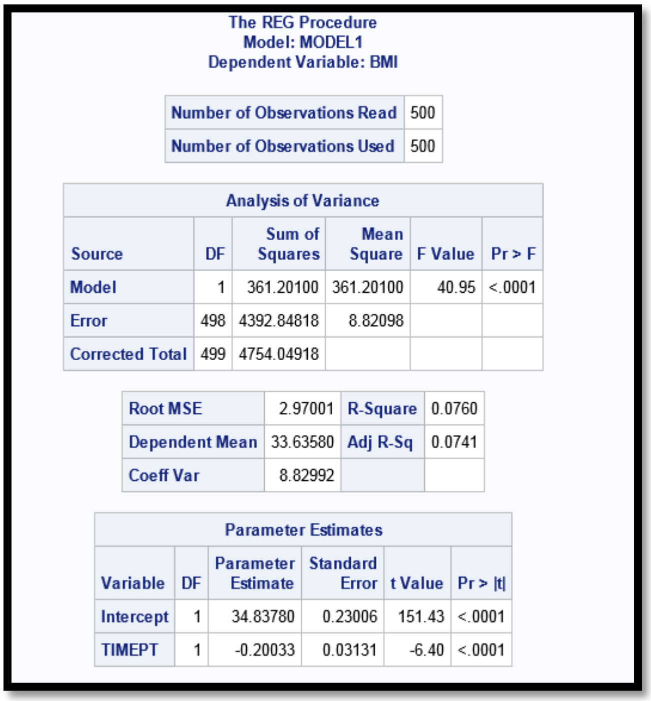
Figure 1. Spaghetti Plot of BMI from Weight Loss Trial

MODELING CONTINUOUS LONGITUDINAL DATA

Now that we have visualized the raw data on our study subjects, let's start thinking about analyzing the data. To start we will we will examine the association between our X value (TIMEPT) and our continuous outcome Y value (BMI) using PROC MIXED, similar to how we would use PROC REG. However, PROC MIXED uses maximum likelihood estimation instead of ordinary least squares to obtain estimates. What is neat about PROC MIXED is that you can use a CLASS statement for categorical variables rather than creating dummy coded variables as you would need for PROC REG. The following is a comparison of PROC REG and PROC MIXED syntax for a simple linear regression model.

```
PROC REG DATA = BMILONG;
      MODEL BMI = TIMEPT;
      RUN;

PROC MIXED DATA = BMILONG;
      MODEL BMI = TIMEPT / SOLUTION;
      RUN;
```

Note that the only difference between these two examples is the PROC declaration of MIXED versus REG, and in PROC MIXED we specify the SOLUTION option in the model statement. By default, PROC MIXED suppresses the parameter estimates so we use /SOLUTION to make sure we can see the resulting model coefficients.



The REG Procedure
Model: MODEL1
Dependent Variable: BMI

Number of Observations Read	500
Number of Observations Used	500

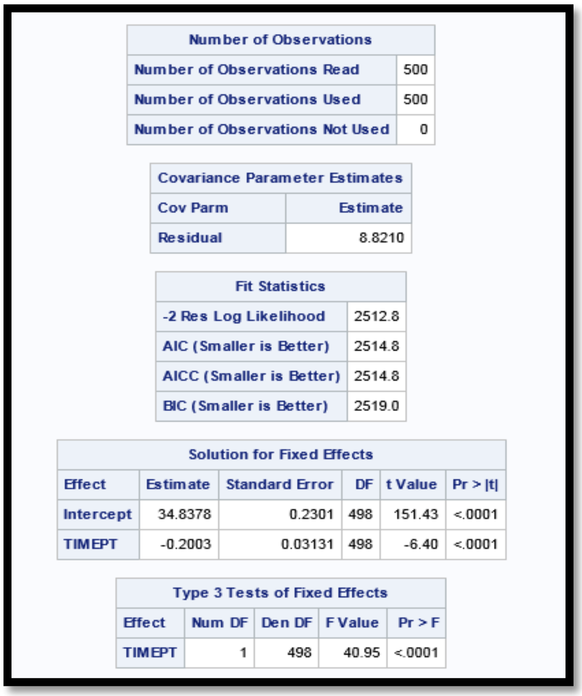
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	361.20100	361.20100	40.95	<.0001
Error	498	4392.84818	8.82098		
Corrected Total	499	4754.04918			

Root MSE	2.97001	R-Square	0.0760
Dependent Mean	33.63580	Adj R-Sq	0.0741
Coeff Var	8.82992		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	34.83780	0.23006	151.43	<.0001
TIMEPT	1	-0.20033	0.03131	-6.40	<.0001



Number of Observations	
Number of Observations Read	500
Number of Observations Used	500
Number of Observations Not Used	0

Covariance Parameter Estimates

Cov Parm	Estimate
Residual	8.8210

Fit Statistics	
-2 Res Log Likelihood	2512.8
AIC (Smaller is Better)	2514.8
AICC (Smaller is Better)	2514.8
BIC (Smaller is Better)	2519.0

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	34.8378	0.2301	498	151.43	<.0001
TIMEPT	-0.2003	0.03131	498	-6.40	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
TIMEPT	1	498	40.95	<.0001

A.PROC REG OUTPUT **B.PROC MIXED OUTPUT**

Output 3. Output Comparison of PROC REG with PROC MIXED

Output 3 shows the output from PROC REG on the left and the output of PROC MIXED on the right. You can match up some of the data elements so you get a feel for the layout of the output in PROC MIXED. On both outputs you can see the number of observations read and used. PROC MIXED has an additional line which shows the number of observations not used, i.e. the difference when you subtract number of observations used from number of observations read. PROC REG on the left has Analysis of Variance which PROC MIXED does not, but you if you look carefully at the output of PROC MIXED you can see that some of that information is there. Since we are only testing one continuous covariate in the model, the Analysis of Variance F test will have the same results as the Type 3 Test. Therefore, we can see the 40.95 from the F value of the Analysis of Variance is identical to the F Value in the Type 3 Tests of Fixed Effects on the bottom of the PROC MIXED output. Also, the Covariance Parameter Estimate of 8.8210 from the PROC MIXED output on the right is approximately equal to the Mean Square Error of 8.82098 from the PROC REG output on the left. PROC MIXED does not have R-Square information since it uses maximum likelihood for the modeling as opposed to ordinary least squares regression. So instead of R-square information, PROC MIXED output uses AIC and other fit statistics to explain model fit. AIC is the Akaike Information Criterion and it is an estimate of the log likelihood corrected for the number of parameters in the model. More information on this statistic can be found in SAS documentation. AIC doesn't have a nice interpretation like R-Square which is used to describe the amount of variation in response explained by the model. However, AIC can be used for model selection which we will describe later in this HOW. Finally, note that the output of parameter estimates and the solution for fixed effects are essentially the same. Both of these are the estimated model coefficients.

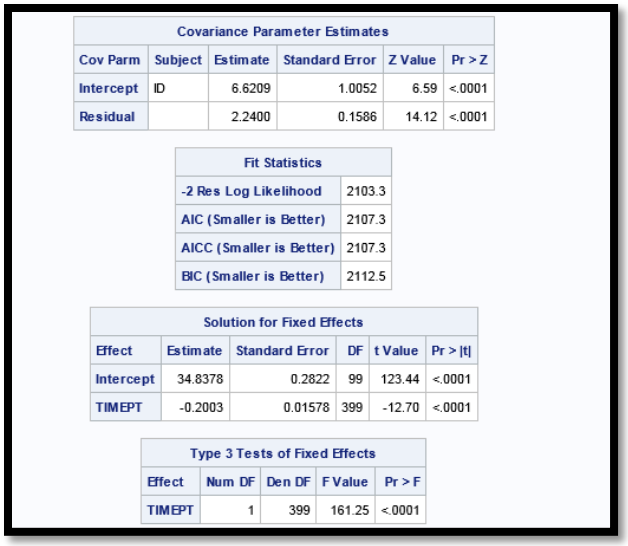
Now that we have compared PROC REG to PROC MIXED we will go forward with PROC MIXED and use it correctly for a repeated measures analysis by correcting for the violation of independence assumption.

MODELING THE RANDOM EFFECTS

The modeling of longitudinal data can be split into modeling FIXED effects and RANDOM effects. Modeling FIXED effects is similar to running analyses with PROC REG. This is the modeling of the dependent response using the covariates. Now, we will introduce the modeling of RANDOM effects. Don't be afraid of this new term RANDOM effects. This is just the partitioning of the error term. What I have often encountered is that the RANDOM effects statement is what often intimidates people from wanting to perform longitudinal data analyses. It helps to think that we can continue to evaluate covariates or FIXED effects as we ordinarily would and we are just now considering how to partition the error. Specifically, the error is being partitioned into a subject specific error and a random error. This makes sense since our independence assumption is violated and that is why we are using longitudinal analysis. There are two ways to model the random effects in longitudinal analyses. We can look at a whether or not the subject specific intercepts and slopes are different from the overall cohort with RANDOM statements or we can look at the way data across time points are correlated using REPEATED statement. Let's start with a simple comparison of modeling RANDOM effects with RANDOM and REPEATED statements.

```
PROC MIXED DATA = BMILONG METHOD = REML COVTEST;
      CLASS ID;
      MODEL BMI = TIMEPT / SOLUTION;
      RANDOM INTERCEPT / SUBJECT = ID;
      RUN;

PROC MIXED DATA = BMILONG METHOD = REML COVTEST;
      CLASS ID;
      MODEL BMI = TIMEPT / SOLUTION;
      REPEATED / SUBJECT = ID TYPE = CS;
      RUN;
```

There are several things we added to these statements from the previous simple PROC MIXED declaration we made when we were comparing it to PROC REG. First, note that in the PROC MIXED statement, we added METHOD=REML and COVTEST. Whenever we are evaluating RANDOM effects, we want to be sure the METHOD is set to REML which is residual (restricted) maximum likelihood. Conversely, we will see that when we model FIXED effects, the METHOD will be set to METHOD=ML or maximum likelihood. Second, specifying the COVTEST option means we want to obtain the significance test of the random effects. This will help us make determinations in specifying the random effects part of the model. The next statement added to PROC MIXED is the CLASS statement where we declare if variables are categorical. Since ID is a numeric variable we want to be sure PROC MIXED understands that ID is categorical variable and not continuous, so it is added to the CLASS statement. Likewise, any other categorical variables we would test should be added to the CLASS statement. Finally, the last statement added is either a RANDOM statement or a REPEATED statement. The RANDOM statement in the first PROC MIXED is testing whether or not there are subject specific intercepts that are significantly different from the overall intercept. The REPEATED statement in the second PROC MIXED is testing whether time points have a compound symmetric pattern, TYPE=CS. Compound symmetric means the correlation of BMIs at any two time points is the same, i.e. the correlation between BASELINE and MONTH3 is the same as the correlation as BASELINE and MONTH6 is the same as the correlation between BASELINE and MONTH12, etc. Note that I use the word correlation to explain the structures, but the output is specified as covariance. Correlation and covariance are synonymous in meaning here and one can obtain covariance from correlation using a basic transformation found in any introductory statistics book. Let us observe the output of modeling the random effects in these two different ways.



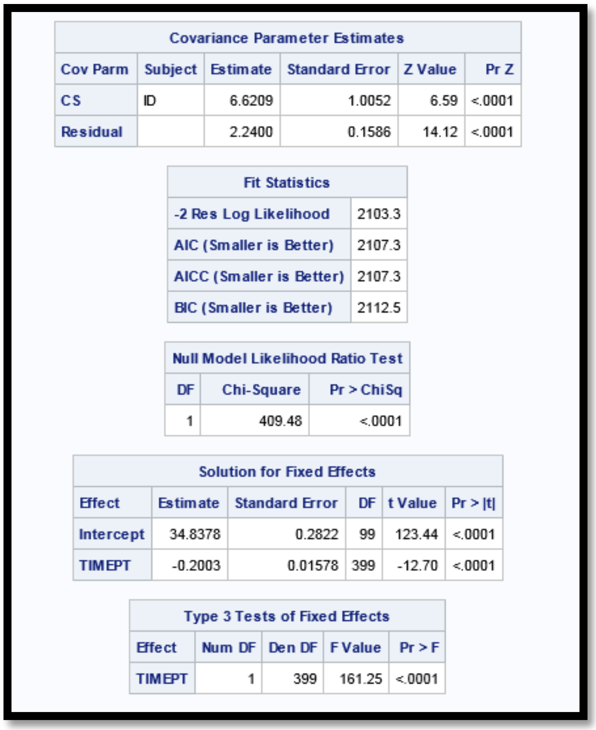
Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
Intercept	ID	6.6209	1.0052	6.59	<.0001
Residual		2.2400	0.1586	14.12	<.0001

Fit Statistics	
-2 Res Log Likelihood	2103.3
AIC (Smaller is Better)	2107.3
AICC (Smaller is Better)	2107.3
BIC (Smaller is Better)	2112.5

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	34.8378	0.2822	99	123.44	<.0001
TIMEPT	-0.2003	0.01578	399	-12.70	<.0001

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
TIMEPT	1	399	161.25	<.0001

A.RANDOM STATEMENT OUTPUT



Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
CS	ID	6.6209	1.0052	6.59	<.0001
Residual		2.2400	0.1586	14.12	<.0001

Fit Statistics	
-2 Res Log Likelihood	2103.3
AIC (Smaller is Better)	2107.3
AICC (Smaller is Better)	2107.3
BIC (Smaller is Better)	2112.5

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
1	409.48	<.0001

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	34.8378	0.2822	99	123.44	<.0001
TIMEPT	-0.2003	0.01578	399	-12.70	<.0001

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
TIMEPT	1	399	161.25	<.0001

B .REPEATED STATEMENT OUTPUT

Output 4. Output Comparison of RANDOM INTERCEPT Statement versus REPEATED TYPE=CS Statement

Looking at Output 4 compares the random effects parameter estimates in a model using the RANDOM statement with RANDOM INTERCEPT specified and a model using the REPEATED statement with option TYPE=CS specified. Looking at the Covariance Parameter Estimates in both outputs on the left and right, we see that the results are identical. Conceptually, they may seem different because with the RANDOM statement we are looking at varying the intercept by subject and in the REPEATED statement we are looking at whether any two time points have the same correlation. However, what this output shows is that although these two sentiments may seem conceptually different, they will yield the same result in the model output. Notice that all numbers in the two outputs are the same. The REPEATED output also has the Null Model Likelihood Ratio Test which tells us whether to model the random effects of the data at all, i.e. whether or not any of the Covariance Parameter Estimates are significant. The significant p-value (<0.05) of the Null Model Likelihood Ratio Test tells us there are significant random effects overall in the model which we can also see in the Covariance Parameter Estimates output.

Now that we know the results are the same, what does it all mean? The Covariance Parameter Estimate on the ID line (Intercept in Output 4A or CS in Output 4B) is the amount of error attributed to the subject. Recall in the previous Output 3 that the overall residual was 8.8210. Now the residual is split into the subject specific error of 6.6209 and the residual or random error of 2.2400. When added together, we get 8.8609, approximately 8.8. The small variation may be attributed to the iterative process and convergence method use by maximum likelihood and random (restricted) maximum likelihood to arrive at these estimates.

The Fit Statistics in the two models (Outputs 4A and 4B) are identical and what is of note is that the fixed effects parameter estimates don't change at all from Output 3. The standard errors and degrees of freedom are different though which may have an impact on significance of fixed effects. This is why it is important to separate out testing of fixed effects and random effects. Focus on testing one aspect of the model at a time, while ignoring the other. I like to fit random effects first and then test covariates in fixed effects but the choice is arbitrary. Also, although the parameter estimates between Output 3 and Output 4 are identical, you may observe slight changes due to iterative nature of PROC MIXED estimation.

We just examined two types of random effects, random intercept model and repeated measures model with compound symmetric covariance structure, and we noted that these had the same output though conceptually they may be different. The random effects model could be further enhanced by testing for a random slope, a random intercept and slope, and evaluating the many covariance structures listed in SAS documentation. In this HOW we will evaluate two more covariance structures as well as the random slope and random intercept and slope and see which one will best fit in our model. In order to compare model fit, we will look at AIC as we vary the random effects specified and then look for the model with the smallest AIC value. Remember that in testing random effects we want to make sure that we are using METHOD=REML. Table 1 tracks many of the different AICs as we change the specified random effects or even write code using ODS commands to make a data set of all AICs. Many of these are also copied and pasted into the following code so we have it for reference later. Note that AICC and BIC may also be used for model selection, but in this HOW we will focus on AIC.

```
PROC MIXED DATA=BMILONG METHOD=REML COVTEST;
      CLASS ID;
      MODEL BMI = TIMEPT/SOLUTION;
      RANDOM INTERCEPT/SUBJECT=ID;      *Random Intercept-- AIC = 2107.3;
      *RANDOM TIMEPT/SUBJECT=ID;        *Random Slope -- AIC = ;
      *RANDOM INTERCEPT TIMEPT/SUBJECT=ID;*Random Intercept and Slope--AIC =;

      *REPEATED /SUBJECT=ID TYPE=CS;   *Compound Symmetry -- AIC = 2107.3;
      *REPEATED /SUBJECT=ID TYPE=AR(1);*Autoregressive -- AIC=?;
      *REPEATED /SUBJECT=ID TYPE=UN;   *Unstructured -- AIC=?;
      RUN;
```

Un-commenting each RANDOM and REPEATED statement one at a time (while commenting the previous one) allows us to fill in the remaining numbers for the AICs in these various models with different specified random effects. Table 1 illustrates the definition of each one of these specified random effects and the resulting AIC obtained from running each model.

Random Effect Statement	Type	Description

`RANDOM INTERCEPT/SUBJECT=ID;`
`* AIC = 2107.3`
	Random Intercept	Intercept varies by subject

`RANDOM TIMEPT/SUBJECT=ID;`
`* AIC = 2092.5`
	Random Slope	Slope varies by subject

`RANDOM INTERCEPT TIMEPT/SUBJECT=ID;`
`* AIC = 1801.1`
	Random Intercept and Slope	Intercept and Slope vary by subject

`REPEATED /SUBJECT=ID TYPE=CS;`
`* AIC = 2107.3`
	Repeated measures with compound symmetric covariance	Any two time points have the **SAME** correlation, r.

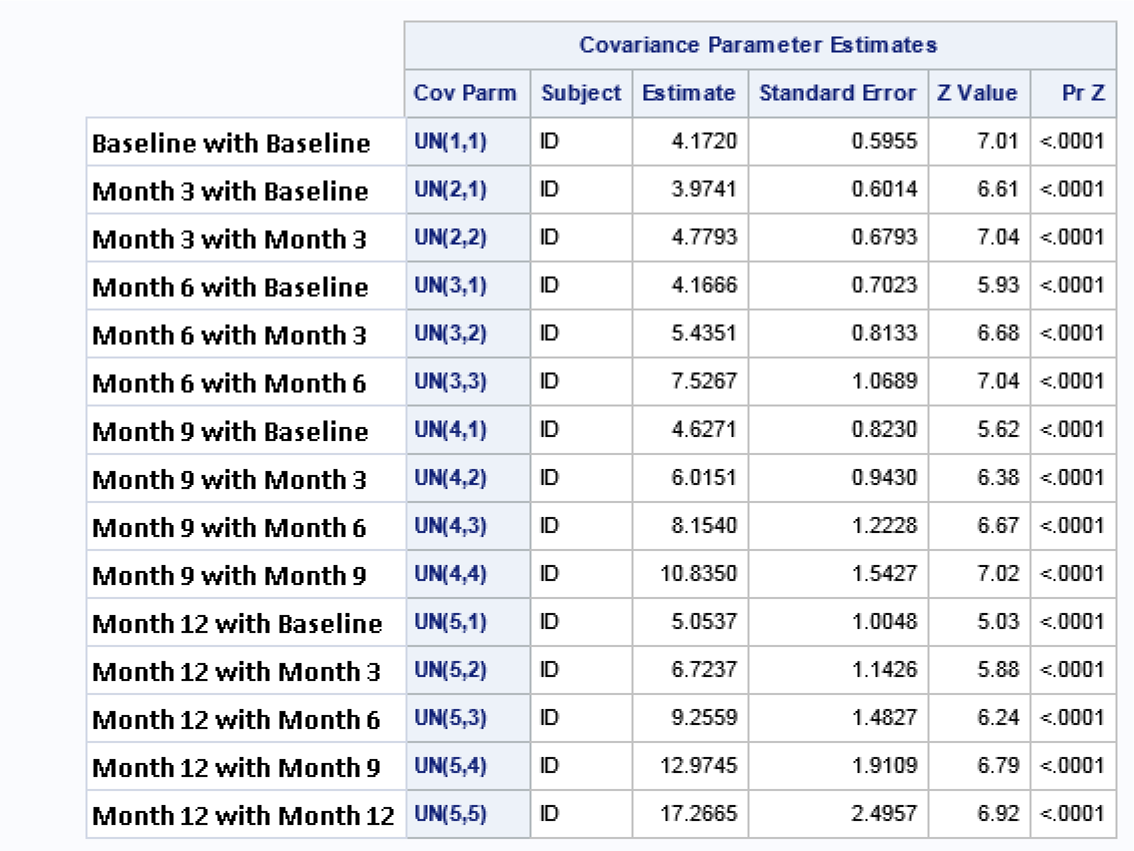
`REPEATED /SUBJECT=ID TYPE=AR(1);`
`* AIC = 1843.4`
	Repeated measures with compound symmetric covariance	The correlation of two time points changes with the interval between observations, i.e. Baseline and month 3 = r, baseline and month 6 = r^2, baseline and month 9 = r^3. Note that as interval increases, absolute correlation decreases since |r|<1.

`REPEATED /SUBJECT=ID TYPE=UN;`
`* AIC = 1724.2`
	Repeated measures with unstructured covariance	Any two time points have **DIFFERENT** correlations.

Table 1. Random Effects Specification and Meaning

We find that after running these various models that the last model with a REPEATED statement using TYPE=UN is the random effects specification which has the lowest AIC at 1724.2. Per Table 1, we find that the model specification that best fits for our data is to have any two time points have a different correlation. So the correlation for BASELINE with MONTH 3 is different from BASELINE with MONTH 6 is different from MONTH 3 with MONTH 6, etc. This can be seen by looking at the Covariance Parameter Estimates in the annotated Output 5. You are welcome to evaluate other covariance structures, but for the convenience of time and paper length we only evaluate these few here.

```
PROC MIXED DATA=BMILONG METHOD=REML COVTEST;
      CLASS ID;
      MODEL BMI = TIMEPT/SOLUTION;
      REPEATED /SUBJECT=ID TYPE=UN;  *Unstructured -- AIC=1724.2;
      RUN;
```



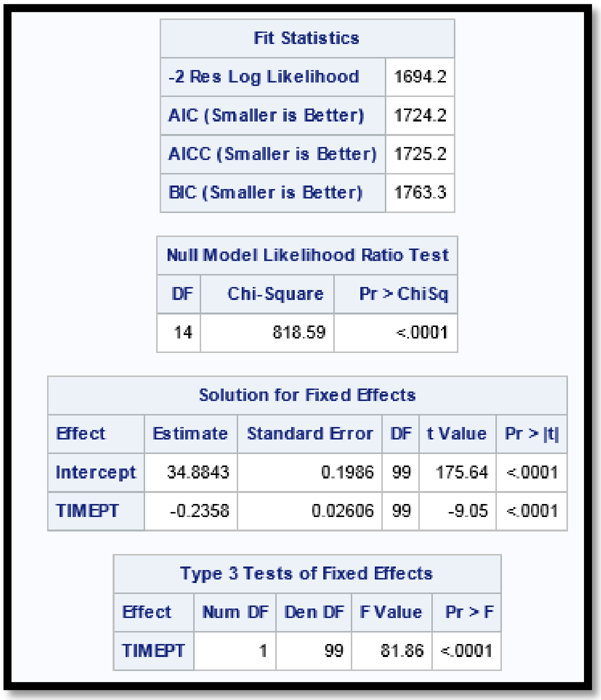
	Covariance Parameter Estimates					
	Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Baseline with Baseline	UN(1,1)	ID	4.1720	0.5955	7.01	<.0001
Month 3 with Baseline	UN(2,1)	ID	3.9741	0.6014	6.61	<.0001
Month 3 with Month 3	UN(2,2)	ID	4.7793	0.6793	7.04	<.0001
Month 6 with Baseline	UN(3,1)	ID	4.1666	0.7023	5.93	<.0001
Month 6 with Month 3	UN(3,2)	ID	5.4351	0.8133	6.68	<.0001
Month 6 with Month 6	UN(3,3)	ID	7.5267	1.0689	7.04	<.0001
Month 9 with Baseline	UN(4,1)	ID	4.6271	0.8230	5.62	<.0001
Month 9 with Month 3	UN(4,2)	ID	6.0151	0.9430	6.38	<.0001
Month 9 with Month 6	UN(4,3)	ID	8.1540	1.2228	6.67	<.0001
Month 9 with Month 9	UN(4,4)	ID	10.8350	1.5427	7.02	<.0001
Month 12 with Baseline	UN(5,1)	ID	5.0537	1.0048	5.03	<.0001
Month 12 with Month 3	UN(5,2)	ID	6.7237	1.1426	5.88	<.0001
Month 12 with Month 6	UN(5,3)	ID	9.2559	1.4827	6.24	<.0001
Month 12 with Month 9	UN(5,4)	ID	12.9745	1.9109	6.79	<.0001
Month 12 with Month 12	UN(5,5)	ID	17.2665	2.4957	6.92	<.0001

Output 5. Output of Covariance Parameter Estimates using REPEATED with TYPE=UN unstructured covariance

MODELING THE FIXED EFFECTS

Now that we have found the best model for the random effects which accounts for the non-independent nature of our data, we can now focus on testing the fixed effects. Recall that we would test these just as we would in a linear regression model in PROC REG. Also, now that we are evaluating fixed effects we need to change our method to METHOD=ML, maximum likelihood. Note that this change alone, even without adding the fixed effects, will slightly change the model results.

```
PROC MIXED DATA=BMILONG METHOD=REML COVTEST;
      CLASS ID;
      MODEL BMI = TIMEPT/SOLUTION;
      REPEATED /SUBJECT=ID TYPE=UN;
      RUN;

PROC MIXED DATA=BMILONG METHOD=ML COVTEST;
      CLASS ID;
      MODEL BMI = TIMEPT/SOLUTION;
      REPEATED /SUBJECT=ID TYPE=UN;
      RUN;
```



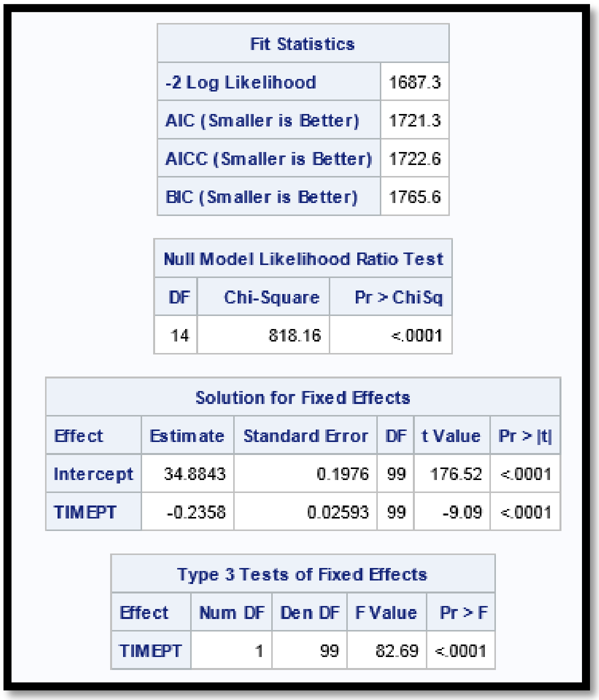
Fit Statistics	
-2 Res Log Likelihood	1694.2
AIC (Smaller is Better)	1724.2
AICC (Smaller is Better)	1725.2
BIC (Smaller is Better)	1763.3

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
14	818.59	<.0001

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	34.8843	0.1986	99	175.64	<.0001
TIMEPT	-0.2358	0.02606	99	-9.05	<.0001

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
TIMEPT	1	99	81.86	<.0001

A .METHOD=REML



Fit Statistics	
-2 Log Likelihood	1687.3
AIC (Smaller is Better)	1721.3
AICC (Smaller is Better)	1722.6
BIC (Smaller is Better)	1765.6

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
14	818.16	<.0001

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	34.8843	0.1976	99	176.52	<.0001
TIMEPT	-0.2358	0.02593	99	-9.09	<.0001

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
TIMEPT	1	99	82.69	<.0001

B .METHOD=ML

Output 6. Output Comparison of METHOD=REML and METHOD=ML

Output 6 shows the result of making a simple change in the METHOD statement from using METHOD=REML in the random effects analyses to METHOD=ML which we will now use for fixed effects analyses. Nearly everything changes when using this computational method; the AIC and the Fixed Effects standard errors are different as shown here and the covariance parameter estimates also changes though not shown here. It is important to note that not only do we want to use METHOD=ML for evaluating FIXED effects but we also want to use METHOD=ML for reporting our final estimates and generating any model predicted values. We demonstrate in Output 6 that altering the METHOD will change the output so in the future you are wary of which METHOD you are using in modeling.

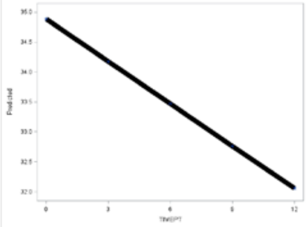
Just as we evaluated multiple random effects model specifications, let's also examine different specifications for the fixed effects. What is nice about PROC MIXED is the ability to use a CLASS statement to specify categorical variables. We previously specified ID as categorical, and now we can also add the GENDER (female=1 versus male=0) and TREAT (treatment =1 versus placebo = 0) variables to the CLASS statement. In our simulation, GENDER and TREAT are dummy coded 0-1 so placing them in the CLASS statement is not required. However, in the future you may receive a data set that has character information or coded 1 and 2 and these types of variables would definitely need to be included in the CLASS statement. Therefore, for the sake of exercise we will specify categorical variables in the CLASS statement. Another great aspect of PROC MIXED is the ability to test interactions directly in the model statement without having to use a DATA step to create additional interaction based variables. In the following code, test different FIXED EFFECTS models to see which one has the lowest AIC just as we did previously for the RANDOM effects modeling.

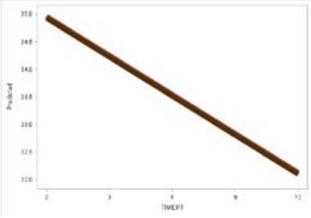
We will also make an addition to this code in order to get predicted values for plotting: OUTPRED=P creates a data set called P with the model predicted values for the fixed effects. After the following PROC MIXED code, we use PROC SORT with a NODUPKEY option on the P data set of model predicted values to get unique predicted values for each of the four gender/treatment combination at each time point output to the PND data set. The SGPLOT code creates a graph of the model estimated BMI (predicted values) against time point to allow for visualization of the model results.

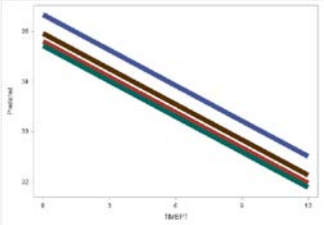
```
PROC MIXED DATA=BMILONG METHOD=ML COVTEST;
      CLASS ID GENDER TREAT;
      MODEL BMI = TIMEPT/SOLUTION OUTPRED=P;*Linear Time – AIC =1721.3;

      *MODEL BMI = TIMEPT GENDER TREAT/SOLUTION OUTPRED=P;
                  *add Main Effects (Different Intercepts);
      *MODEL BMI = TIMEPT GENDER TREAT GENDER*TREAT
                        /SOLUTION OUTPRED=P;
                  *add Interaction (Different Intercepts);
      *MODEL BMI = TIMEPT GENDER TREAT GENDER*TREAT GENDER*TIMEPT
                        TREAT*TIMEPT GENDER*TREAT*TIMEPT
                        /SOLUTION OUTPRED=P;
                        *add Interactions with Time
                        (Different Intercepts and Slopes);
      *MODEL BMI = TIMEPT GENDER TREAT GENDER*TREAT GENDER*TIMEPT
                        TREAT*TIMEPT GENDER*TREAT*TIMEPT
                        TIMEPT*TIMEPT GENDER*TIMEPT*TIMEPT
                        TREAT*TIMEPT*TIMEPT GENDER*TREAT*TIMEPT*TIMEPT
                        /SOLUTION OUTPRED=P;
                                    * add Quadratic Effect and Interactions
                                          (Different Intercepts and Slopes);
      REPEATED /SUBJECT=ID TYPE=UN;
      RUN;

PROC SORT DATA=P OUT= PND NODUPKEY; BY GENDER TREAT TIMEPT; RUN;

PROC SGPLOT DATA=PND;
SERIES X=TIMEPT Y=PRED / GROUP=GROUP LINEATTRS=(PATTERN=1 THICKNESS=10);
XAXIS LABEL ='Month' values=(0 to 12 by 3);
YAXIS LABEL ='Model Estimated BMI';
      RUN;
```

Table 2 shows the results of comparing these five models, the interpretation of each model statement, and the visualization from graphing the predicted value by time point. In Table 2, we also list under each model statement the AIC from the PROC MIXED output. In comparing the AICs across the five different models, we see that the best fit model, the one with lowest AIC, is the final model examined which test quadratic effects. This exercise of testing all the models in this sequence is however quite useful and recommended. In the first model we test a simple linear regression of BMI over time and see that the slope is significant. When we add gender and treatment in the second model and then the interaction of treatment and gender in the third model, we see that the AIC is getting higher which means a worse fit. Additionally, the lines are all very close to each other in the interaction model visualization and if we examine the p-values in the Type 3 test in the output, we would see that the gender, treatment, and their interaction are all non significant. Some analysts might stop here and drop gender and treatment from their evaluation since they were not significant. However, when we look at the fourth model which contains the interactions of gender, treatment, and the gender/treatment interaction with time, we find that the interactions with time are significant. Furthermore, when we test a quadratic time effect and interactions in the final model, this turns out to be significant as well, and the final model that accounts for all these effects and interactions has the best fit. In this data simulation, the end result of the quadratic interaction model being best was known by design, but this modeling exercise teaches a very good point. First, even though main effects are not significant in the model like GENDER and TREAT, they should still be tested to see if they impact the slopes and test interactions of main effects with time. Second, don’t assume that slopes are linear, test to see if a quadratic, cubic, or even using log transformed slope would create a fit model.

Model Statement and AIC	Meaning	Visualization

MODEL BMI = TIMEPT
/SOLUTION OUTPRED=P;
*AIC = 1721.3;
	• Simple Linear Regression.
	• BMI decreases linearly with time.
		

MODEL BMI = TIMEPT GENDER
TREAT/SOLUTION OUTPRED=P;
*AIC = 1725.3;
	• Linear Regression with Gender and Treatment Effect
	• Four parallel lines for each gender and treatment combination and all have BMI decreasing with time.
	• Change in intercept by gender and treatment are constant.
	• No significant treatment or gender effect overall.
		

MODEL BMI = TIMEPT GENDER
TREAT GENDER*TREAT
/SOLUTION OUTPRED=P;
*AIC = 1726.3;
	• Linear Regression with Gender and Treatment Effect and their interaction.
	• Four parallel lines for each gender and treatment combination and all have BMI decreasing with time.
	• No significant treatment or gender effect or interaction effect.
	• Highest line is MALE, PLACEBO (Blue), then FEMALE, TREAT (brown),
	then MALE, TREAT (red),
	then FEMALE, PLACEBO (green)
		

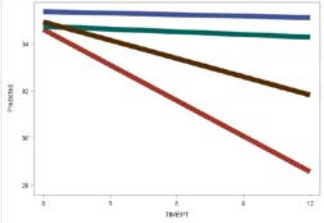
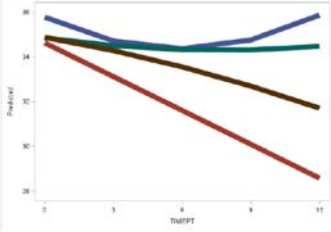
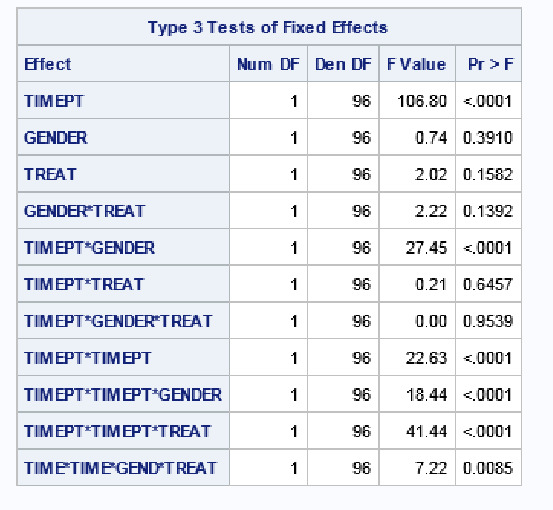
MODEL BMI = TIMEPT GENDER TREAT GENDER*TREAT GENDER*TIMEPT TREAT*TIMEPT GENDER*TREAT*TIMEPT /SOLUTION OUTPRED=P; *AIC =1669.6;	• Linear Regression with Gender and Treatment Effect and Interaction and Interactions with Time • Four lines with different intercepts and slopes. • Significant interactions with time indicating different slopes. • PLACEBO lines are flat and TREAT lines are decreasing	
MODEL BMI = TIMEPT GENDER TREAT GENDER*TREAT GENDER*TIMEPT TREAT*TIMEPT GENDER*TREAT*TIMEPT TIMEPT*TIMEPT GENDER*TIMEPT*TIMEPT TREAT*TIMEPT*TIMEPT GENDER*TREAT*TIMEPT*TIMEPT /SOLUTION OUTPRED=P; *AIC=1620.3;	• Add quadratic effect to all lines • Four different line trends	

Table 2. Fixed Effects Model Statements and Definitions

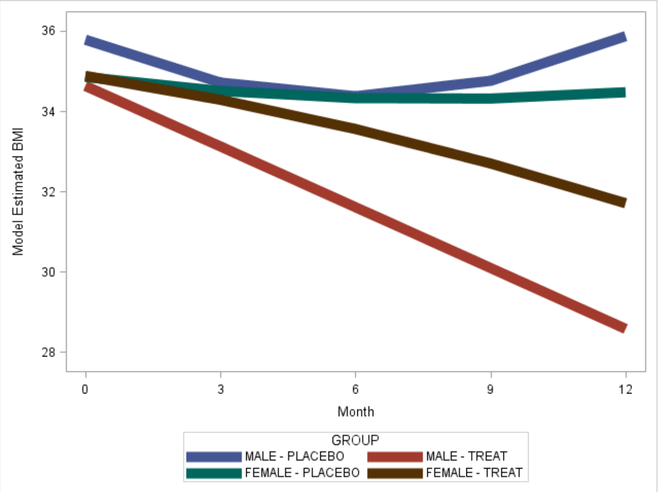
The code for the best model with quadratic effects and interactions is shown below including selected output and a plot of the predicted value from this best model as shown in Output 7.

```
PROC MIXED DATA=BMILONG METHOD=ML COVTEST;
      CLASS ID GENDER TREAT;
      MODEL BMI = TIMEPT GENDER TREAT GENDER*TREAT GENDER*TIMEPT
                        TREAT*TIMEPT GENDER*TREAT*TIMEPT
                        TIMEPT*TIMEPT GENDER*TIMEPT*TIMEPT
                        TREAT*TIMEPT*TIMEPT GENDER*TREAT*TIMEPT*TIMEPT
             /SOLUTION OUTPRED=P; *Quadratic Time + Interactions
                        -- AIC =1620.3;
      REPEATED /SUBJECT=ID TYPE=UN;
      RUN;

PROC SORT DATA=p OUT=pnd NODUPKEY; BY GENDER TREAT TIMEPT; RUN;

PROC SGPLOT DATA=PND;
SERIES X=TIMEPT Y=PRED / GROUP=GROUP LINEATTRS=(PATTERN=1 THICKNESS=10);
XAXIS LABEL='MONTH' VALUES=(0 TO 12 BY 3);
YAXIS LABEL ='Model Estimated BMI';
      RUN;
```



Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
TIMEPT	1	96	106.80	<.0001
GENDER	1	96	0.74	0.3910
TREAT	1	96	2.02	0.1582
GENDER*TREAT	1	96	2.22	0.1392
TIMEPT*GENDER	1	96	27.45	<.0001
TIMEPT*TREAT	1	96	0.21	0.6457
TIMEPT*GENDER*TREAT	1	96	0.00	0.9539
TIMEPT*TIMEPT	1	96	22.63	<.0001
TIMEPT*TIMEPT*GENDER	1	96	18.44	<.0001
TIMEPT*TIMEPT*TREAT	1	96	41.44	<.0001
TIME*TIME*GEND*TREAT	1	96	7.22	0.0085



Output 7. Selected Output from PROC MIXED and PROC SGPLOT using Quadratic Model with Interactions

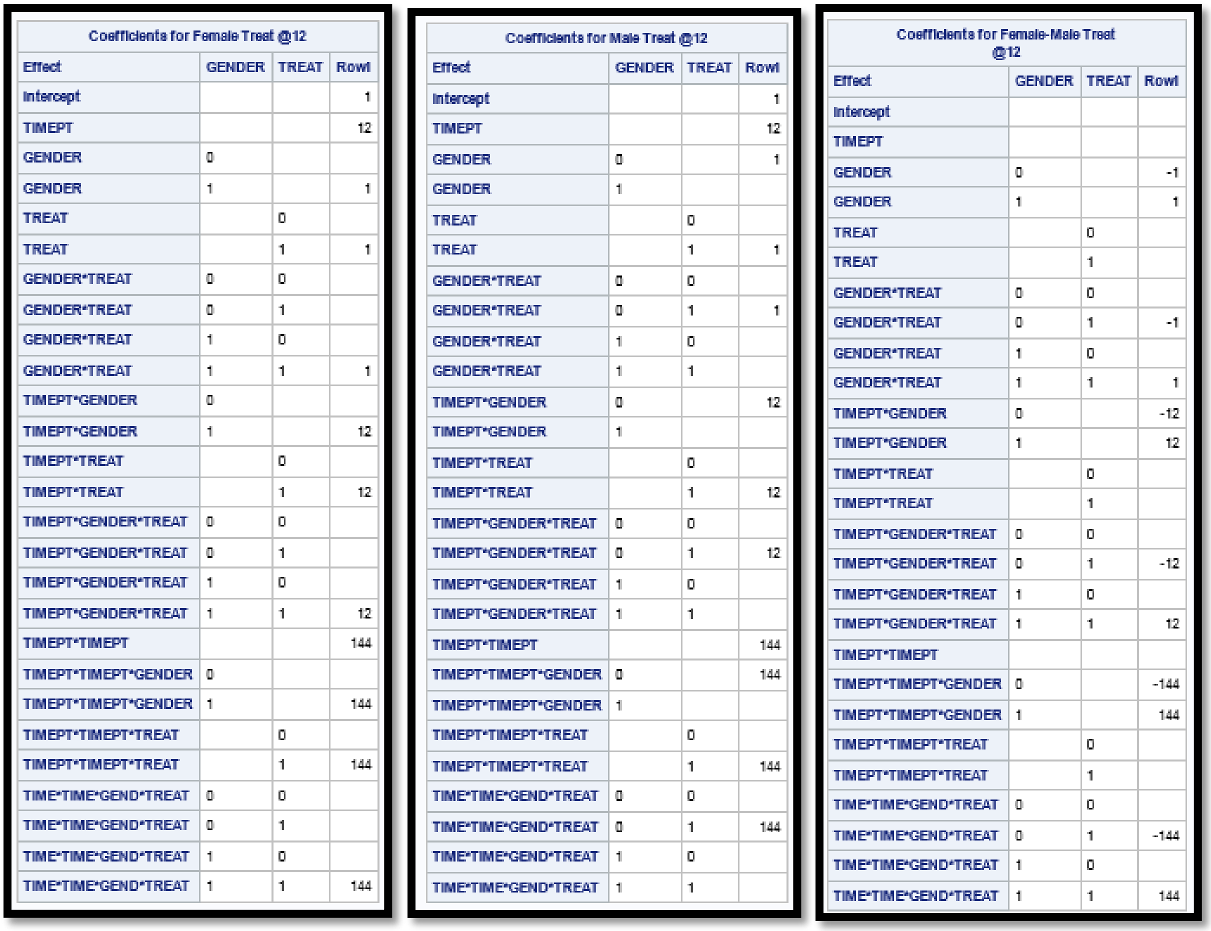
In Output 7, we can see from the Type 3 effects that all the quadratic interactions are significant in the model while some of the other effects are not. By hierarchical principal, it is okay to ignore the non-significance of the main effects if the terms involved in interactions are significant. From the plot we see that as we designed the Male Placebo group (blue line) has a small quadratic change in BMI over time and the female placebo green line is nearly flat. The men on treatment have a linearly decreasing BMI (red line) and the rate at which their BMI decreases is faster than the rate of decrease for females on treatment (brown line).

GETTING MODEL ESTIMATES

Now that we have successfully obtained a well fit model, suppose the investigator comes back and asks 'Is there a significant difference in the slopes between men and women? At what time point does that difference in BMI between men and women become significant?' We couldn’t really get this information from the Type 3 estimates and typically it isn't enough to look at the coefficients, predicted values, or graphs. To answer these questions we will add ESTIMATE statements to our PROC MIXED code. ESTIMATE statements will generate linear combinations of the coefficients in order to get desired predicted values or desired differences in predicted values. This is analogous to plugging in values to the equation to obtain the predicted values. The best way to start writing ESTIMATE statements is to try to regenerate known predicted values and then modify them to suit your needs. Below a few ESTIMATE statements are listed and the results are shown in the output that follows. The E option at the end of each ESTIMATE statement is very useful in showing which values were specified, especially for CLASS variables. The SAS Support Documentation contains more information about all the nuisances of using ESTIMATE statements.

```
PROC MIXED DATA=BMILONG METHOD=ML COVTEST;
      CLASS ID GENDER TREAT;
      MODEL BMI = TIMEPT GENDER TREAT GENDER*TREAT GENDER*TIMEPT
                        TREAT*TIMEPT GENDER*TREAT*TIMEPT
                        TIMEPT*TIMEPT GENDER*TIMEPT*TIMEPT
                        TREAT*TIMEPT*TIMEPT GENDER*TREAT*TIMEPT*TIMEPT
             /SOLUTION OUTPRED=P; *Quadratic Time + Interactions
                        -- AIC =1620.3;
      REPEATED /SUBJECT=ID TYPE=UN;

      ESTIMATE 'Female Treat @12'  INTERCEPT 1 TIMEPT 12 GENDER 0 1
            TREAT 0 1 GENDER*TREAT 0 0 0 1 GENDER*TIMEPT 0 12
                        TREAT*TIMEPT 0 12 GENDER*TREAT*TIMEPT  0 0 0 12
                        TIMEPT*TIMEPT 144  GENDER*TIMEPT*TIMEPT 0 144
      TREAT*TIMEPT*TIMEPT 0 144
                        GENDER*TREAT*TIMEPT*TIMEPT 0 0 0 144/E;

      ESTIMATE 'Male Treat @12'  INTERCEPT 1 TIMEPT 12 GENDER 1 0
      TREAT 0 1 GENDER*TREAT 0 1 0 0 GENDER*TIMEPT 12 0
                        TREAT*TIMEPT 0 12 GENDER*TREAT*TIMEPT 0 12  0 0
                        TIMEPT*TIMEPT 144  GENDER*TIMEPT*TIMEPT 144 0
      TREAT*TIMEPT*TIMEPT 0 144
                        GENDER*TREAT*TIMEPT*TIMEPT 0 144 0  0/E;


      ESTIMATE 'Female-Male Treat @12'   GENDER -1 1
                        GENDER*TREAT 0 -1 0  1 GENDER*TIMEPT -12 12
                         GENDER*TREAT*TIMEPT  0 -12 0 12
                         GENDER*TIMEPT*TIMEPT -144 144
                        GENDER*TREAT*TIMEPT*TIMEPT 0 -144 0 144/E;
      RUN;


PROC SORT DATA=P OUT=PND NODUPKEY; BY GENDER TREAT TIMEPT; RUN;
PROC PRINT DATA=PND;
WHERE TREAT=1 AND TIMEPT=12;
RUN;
```

The preceding code reproduces estimates at TIMEPT=12 or 12 months for females and males on treatment and then estimates the difference between the females – males. Notice how fewer terms are specified in the Female-Male Treat @12 ESTIMATE statement. This is because if we take the second ESTIMATE statement and subtract the first ESTIMATE statement, the terms in common will zero out or drop. The following Outputs 8, 9 and 10 show the results of using these ESTIMATE statements.



Coefficients for Female Treat @12

Effect	GENDER	TREAT	Row1
Intercept			1
TIMEPT			12
GENDER	0		
GENDER	1		1
TREAT		0	
TREAT		1	1
GENDER*TREAT	0	0	
GENDER*TREAT	0	1	
GENDER*TREAT	1	0	
GENDER*TREAT	1	1	1
TIMEPT*GENDER	0		
TIMEPT*GENDER	1		12
TIMEPT*TREAT		0	
TIMEPT*TREAT		1	12
TIMEPT*GENDER*TREAT	0	0	
TIMEPT*GENDER*TREAT	0	1	
TIMEPT*GENDER*TREAT	1	0	
TIMEPT*GENDER*TREAT	1	1	12
TIMEPT*TIMEPT			144
TIMEPT*TIMEPT*GENDER	0		
TIMEPT*TIMEPT*GENDER	1		144
TIMEPT*TIMEPT*TREAT		0	
TIMEPT*TIMEPT*TREAT		1	144
TIME*TIME*GEND*TREAT	0	0	
TIME*TIME*GEND*TREAT	0	1	
TIME*TIME*GEND*TREAT	1	0	
TIME*TIME*GEND*TREAT	1	1	144

Coefficients for Male Treat @12

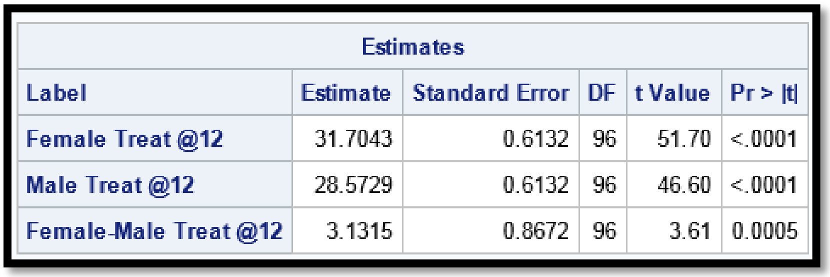
Effect	GENDER	TREAT	Row1
Intercept			1
TIMEPT			12
GENDER	0		1
GENDER	1		
TREAT		0	
TREAT		1	1
GENDER*TREAT	0	0	
GENDER*TREAT	0	1	1
GENDER*TREAT	1	0	
GENDER*TREAT	1	1	
TIMEPT*GENDER	0		12
TIMEPT*GENDER	1		
TIMEPT*TREAT		0	
TIMEPT*TREAT		1	12
TIMEPT*GENDER*TREAT	0	0	
TIMEPT*GENDER*TREAT	0	1	12
TIMEPT*GENDER*TREAT	1	0	
TIMEPT*GENDER*TREAT	1	1	
TIMEPT*TIMEPT			144
TIMEPT*TIMEPT*GENDER	0		144
TIMEPT*TIMEPT*GENDER	1		
TIMEPT*TIMEPT*TREAT		0	
TIMEPT*TIMEPT*TREAT		1	144
TIME*TIME*GEND*TREAT	0	0	
TIME*TIME*GEND*TREAT	0	1	144
TIME*TIME*GEND*TREAT	1	0	
TIME*TIME*GEND*TREAT	1	1	

Coefficients for Female-Male Treat @12

Effect	GENDER	TREAT	Row1
Intercept			
TIMEPT			
GENDER	0		-1
GENDER	1		1
TREAT		0	
TREAT		1	
GENDER*TREAT	0	0	
GENDER*TREAT	0	1	-1
GENDER*TREAT	1	0	
GENDER*TREAT	1	1	1
TIMEPT*GENDER	0		-12
TIMEPT*GENDER	1		12
TIMEPT*TREAT		0	
TIMEPT*TREAT		1	
TIMEPT*GENDER*TREAT	0	0	
TIMEPT*GENDER*TREAT	0	1	-12
TIMEPT*GENDER*TREAT	1	0	
TIMEPT*GENDER*TREAT	1	1	12
TIMEPT*TIMEPT			
TIMEPT*TIMEPT*GENDER	0		-144
TIMEPT*TIMEPT*GENDER	1		144
TIMEPT*TIMEPT*TREAT		0	
TIMEPT*TIMEPT*TREAT		1	
TIME*TIME*GEND*TREAT	0	0	
TIME*TIME*GEND*TREAT	0	1	-144
TIME*TIME*GEND*TREAT	1	0	
TIME*TIME*GEND*TREAT	1	1	144

A. Female on TREAT at TIMEPT=12 B. Male on TREAT at TIMEPT=12 C. TREAT Female-Male at TIMEPT=12

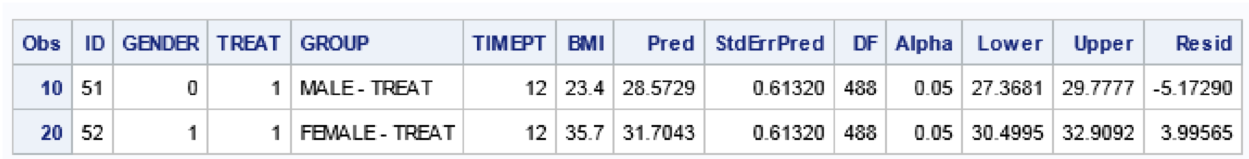
Output 8. Output of ESTIMATE statements with E Option in PROC MIXED



Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Female Treat @12	31.7043	0.6132	96	51.70	<.0001
Male Treat @12	28.5729	0.6132	96	46.60	<.0001
Female-Male Treat @12	3.1315	0.8672	96	3.61	0.0005

Output 9. Output of Estimate Results from ESTIMATE statements in PROC MIXED



Obs	ID	GENDER	TREAT	GROUP	TIMEPT	BMI	Pred	StdErrPred	DF	Alpha	Lower	Upper	Resid
10	51	0	1	MALE - TREAT	12	23.4	28.5729	0.61320	488	0.05	27.3681	29.7777	-5.17290
20	52	1	1	FEMALE - TREAT	12	35.7	31.7043	0.61320	488	0.05	30.4995	32.9092	3.99565

Output 10. Output of PROC PRINT of OUTPRED data set of predicted values

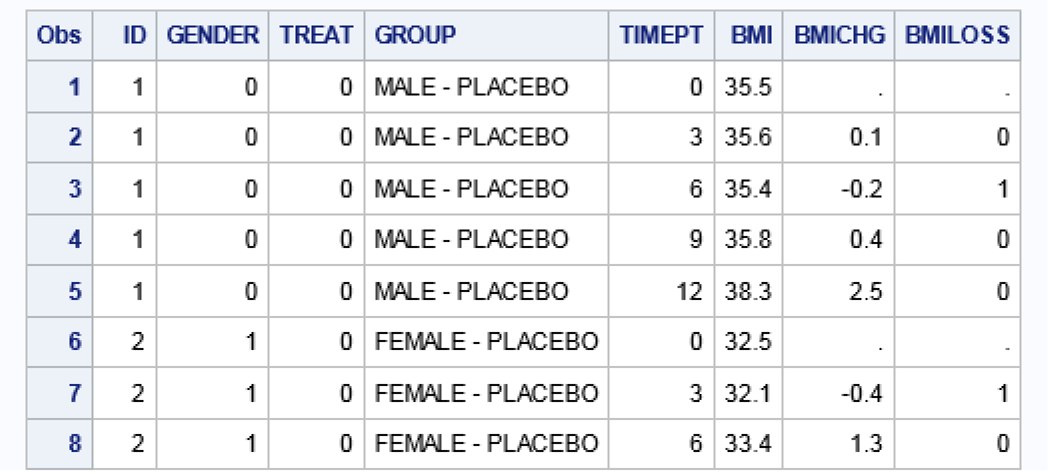
In Output 8, we can see the values we plugged in to the regression equation that produce the predicted values. Using vector subtraction of the Row1 column Output 8B (Male on TREAT at TIME=12) from the Row1 column of Output 8A (Female on TREAT at TIME=12) we can see how we get Output 8C, TREAT Female-Male at TIMEPT=12. We can compare Output 9, the predicted values from the ESTIMATE statements with Output 10 which shows the predicted values for treatment patients at month 12 time point for males and females. The predicted values in PND are identical to estimates we get in Output 9 for males and females which show that we correctly specified the ESTIMATE statements. We could use a calculator and verify that 31.7043 – 28.5729 = 3.1315 which is the difference shown in the third ESTIMATE statement. Additionally, we were able to show the significant difference between the males and females on treatment at 12 months. If we repeat this exercise for other time points 0,3,6,9, we can find the earliest time point (month) when we first estimate a significant difference between male and females and thereby answer our investigator's question.

NON-LINEAR REGRESSION

We have effectively shown so far how to conduct analyses of data when the outcome is continuous and we can use PROC MIXED to perform regression. However, in some cases we may have to deal with outcomes that are not continuous in nature such as count data or dichotomous outcomes. We will briefly show how to analyze such data using either PROC GENMOD or PROC GLIMMIX. The methodology for transforming data and testing random and fixed effects will be similar to the previously discussed sections, though the model specification will change slightly depending on the procedure and analysis.

In the following example we will perform a repeated measures logistic regression analysis on whether or not a patient experienced a loss in BMI, or in other words, a weight loss at any time point. In order to do this we create a new data set called BMILONG1 which will have the difference within each patient of the BMI at a given time point from the previous time point. To do this we first ensure that the data is sorted in proper order by ID and TIMEPT. Then we use the DIF function to create a new variable (BMICHG), which generates the difference in the BMI of the given row from the row above it. Note that there shouldn't be any BMICHG values for TIMEPT=0 since this is baseline. So we add a statement that says if TIMEPT=0 then BMICHG should be set to missing. Next, we create the dichotomous variable BMILOSS which is 1 if BMICHG is negative and 0 if BMICHG is 0 or positive. We only want values when BMICHG isn't missing so we add the condition IF condition at the start of the statement. Output 11 shows the resulting data set BMILONG1 with the additional variables added.

```
PROC SORT data=bmilong; by id TIMEPT;RUN;
DATA BMILONG1;
      SET BMILONG;
       BMICHG=DIF(BMI);
       IF TIMEPT=0 THEN BMICHG=.;
         IF BMICHG NE . THEN BMILOSS = (.<BMICHG<0);
      RUN;
```



Obs	ID	GENDER	TREAT	GROUP	TIMEPT	BMI	BMICHG	BMILOSS
1	1	0	0	MALE - PLACEBO	0	35.5	.	.
2	1	0	0	MALE - PLACEBO	3	35.6	0.1	0
3	1	0	0	MALE - PLACEBO	6	35.4	-0.2	1
4	1	0	0	MALE - PLACEBO	9	35.8	0.4	0
5	1	0	0	MALE - PLACEBO	12	38.3	2.5	0
6	2	1	0	FEMALE - PLACEBO	0	32.5	.	.
7	2	1	0	FEMALE - PLACEBO	3	32.1	-0.4	1
8	2	1	0	FEMALE - PLACEBO	6	33.4	1.3	0

Output 11. BMILONG data set with added columns BMICHG and BMILOSS

Now the data are correctly set-up for a repeated measures logistic regression analysis. We will explore predictors of dichotomous outcomes using PROC GENMOD and PROC GLIMMIX. These two approaches for repeated measures logistic regression analyses give slightly different results. The main difference between GENMOD and GLIMMIX lies in the way that the model estimation is performed. Another main distinction is that only RANDOM statements can be used in GLIMMIX and only REPEATED statements are possible in GENMOD. Consider what we stated previously about the interpretation of RANDOM statements versus REPEATED statements. RANDOM statements help us understand if there are subject level differences for slopes and intercepts. Therefore GLIMMIX models are useful for subject level inferences. REPEATED statements are used to understand how repeated measures are correlated and therefore GENMOD is useful for population based inferences. GENMOD only has a few possible covariance matrix specifications that can be evaluated. Further comparisons between PROC GENMOD and PROC GLIMMIX can be found in the SAS documentation. The following code demonstrates how to model the dichotomous BMILOSS outcome in both PROC GENMOD and PROC GLIMMIX.

```
PROC GLIMMIX DATA=BMILONG1;
      CLASS ID GENDER TREAT;
      MODEL BMILOSS = TIMEPT GENDER TREAT GENDER*TREAT
            TIMEPT*GENDER TIMEPT*TREAT TIMEPT*GENDER*TREAT
            /SOLUTION LINK=LOGIT DIST=BINOMIAL;
      RANDOM INTERCEPT /SUBJECT=ID TYPE=UN;
      RUN;


PROC GENMOD DATA=BMILONG1 DESC;
      CLASS ID GENDER TREAT BMILOSS;
      MODEL BMILOSS = TIMEPT GENDER TREAT GENDER*TREAT
            TIMEPT*GENDER TIMEPT*TREAT TIMEPT*GENDER*TREAT
            / LINK=LOGIT DIST=BINOMIAL;
      REPEATED SUBJECT=ID/ TYPE=CS;
      RUN;
```

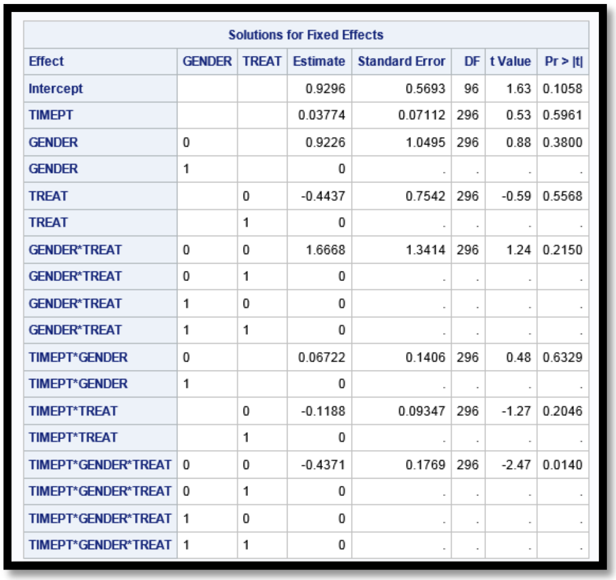
If we evaluate the code for PROC GLIMMIX compared PROC GENMOD line by line, we can see the distinction in the two model specifications.

1. In the PROC statement, GENMOD requires a descending option DESC for the model to correctly assign BMILOSS=1 as the outcome where GLIMMIX does this automatically. There are METHODS which could be varied in PROC GLIMMIX (not PROC GENMOD) that are not covered here, more information can be found in SAS Documentation for PROC GLIMMIX.

2. The model statements in the two PROCs are nearly the same. In PROC GLIMMIX, we specify the SOLUTION option to get the fixed effects parameter estimates but in PROC GENMOD these are output automatically. In both procedures, we specify the LINK=LOGIT and DIST=BINOMIAL which are used for specifying the correct distribution for logistic regression.

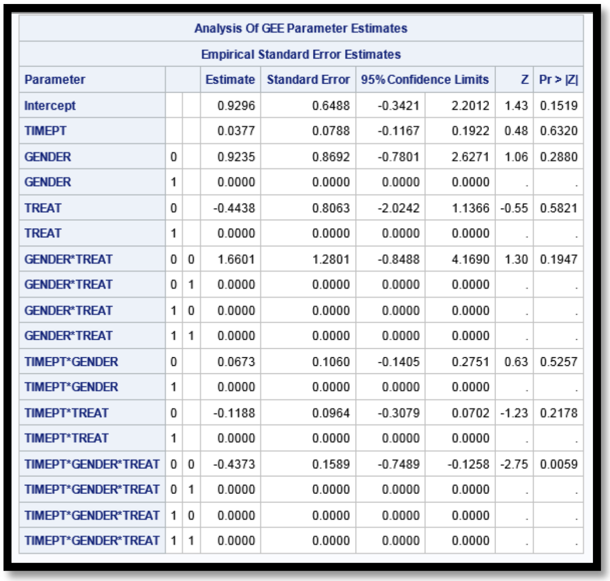
3. In PROC GLIMMIX we specify a RANDOM statement and in PROC GENMOD we specify a REPEATED statement. The SUBJECT=ID specification is listed after the slash (/) in PROC GLIMMIX, just as it was for PROC MIXED. But notice that in PROC GENMOD, we put the SUBJECT=ID before the slash (/). In order to make the results parallel between the GLIMMIX and GENMOD, we specify the RANDOM INTERCEPT statement in PROC GLIMMIX which we know from Output 4 above should be equivalent to the REPEATED with TYPE=CS compound symmetry covariance specified in PROC GENMOD.

Output 12 compares shows the comparison of the output of PROC GLIMMIX and PROC GENMOD.



Solutions for Fixed Effects

Effect	GENDER	TREAT	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			0.9296	0.5693	96	1.63	0.1058
TIMEPT			0.03774	0.07112	296	0.53	0.5961
GENDER	0		0.9226	1.0495	296	0.88	0.3800
GENDER	1		0	.	.	.	.
TREAT		0	-0.4437	0.7542	296	-0.59	0.5568
TREAT		1	0	.	.	.	.
GENDER*TREAT	0	0	1.6668	1.3414	296	1.24	0.2150
GENDER*TREAT	0	1	0	.	.	.	.
GENDER*TREAT	1	0	0	.	.	.	.
GENDER*TREAT	1	1	0	.	.	.	.
TIMEPT*GENDER	0		0.06722	0.1406	296	0.48	0.6329
TIMEPT*GENDER	1		0	.	.	.	.
TIMEPT*TREAT		0	-0.1188	0.09347	296	-1.27	0.2046
TIMEPT*TREAT		1	0	.	.	.	.
TIMEPT*GENDER*TREAT	0	0	-0.4371	0.1769	296	-2.47	0.0140
TIMEPT*GENDER*TREAT	0	1	0	.	.	.	.
TIMEPT*GENDER*TREAT	1	0	0	.	.	.	.
TIMEPT*GENDER*TREAT	1	1	0	.	.	.	.



Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter			Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept			0.9296	0.6488	-0.3421	2.2012	1.43	0.1519
TIMEPT			0.0377	0.0788	-0.1167	0.1922	0.48	0.6320
GENDER	0		0.9235	0.8692	-0.7801	2.6271	1.06	0.2880
GENDER	1		0.0000	0.0000	0.0000	0.0000	.	.
TREAT	0		-0.4438	0.8063	-2.0242	1.1366	-0.55	0.5821
TREAT	1		0.0000	0.0000	0.0000	0.0000	.	.
GENDER*TREAT	0	0	1.6601	1.2801	-0.8488	4.1690	1.30	0.1947
GENDER*TREAT	0	1	0.0000	0.0000	0.0000	0.0000	.	.
GENDER*TREAT	1	0	0.0000	0.0000	0.0000	0.0000	.	.
GENDER*TREAT	1	1	0.0000	0.0000	0.0000	0.0000	.	.
TIMEPT*GENDER	0		0.0673	0.1060	-0.1405	0.2751	0.63	0.5257
TIMEPT*GENDER	1		0.0000	0.0000	0.0000	0.0000	.	.
TIMEPT*TREAT	0		-0.1188	0.0964	-0.3079	0.0702	-1.23	0.2178
TIMEPT*TREAT	1		0.0000	0.0000	0.0000	0.0000	.	.
TIMEPT*GENDER*TREAT	0	0	-0.4373	0.1589	-0.7489	-0.1258	-2.75	0.0059
TIMEPT*GENDER*TREAT	0	1	0.0000	0.0000	0.0000	0.0000	.	.
TIMEPT*GENDER*TREAT	1	0	0.0000	0.0000	0.0000	0.0000	.	.
TIMEPT*GENDER*TREAT	1	1	0.0000	0.0000	0.0000	0.0000	.	.

A. Parameter Estimates GLIMMIX **B. Parameter Estimates GENMOD**

Output 12. Output of Parameter Estimates Comparing GLIMMIX with GENMOD

In viewing Output 12, we can compare the parameter estimates of GLIMMIX and GENMOD and see that all of the estimates are similar. The variation is likely due to the in differences of computation methods in the two procedures. The outputs of random effects in the two models are not easily compared without additional calculation since PROC GLIMMIX outputs covariance estimates and PROC GENMOD outputs correlation estimates. As stated previously, in this workshop covariance can be transformed to correlation simply using a formula found in any introductory statistics book. The two procedures, GENMOD and GLIMMIX, are different in computational method, specification, and purpose. However, when specified in a similar way, these two approaches to repeated measures logistic regression have similar output.

CONCLUSION

This workshop has provided an introductory lesson on longitudinal and repeated measures modeling. We covered how to properly format a data set, analyze continuous and dichotomous outcomes, model random and fixed effects, and graph observed and predicted values using a simple simulated longitudinal data set. We recognize that in reality most data sets encountered won't be this clean and straightforward, but this example BMI data set serves the purpose of emphasizing key repeated measures concepts. We also emphasized longitudinal data analyses with repeated measurements on a subject over time rather than a repeated measures framework without a time aspect such as kids in the same classroom. There are many more repeated measure analyses that can be performed with SAS. The key concept to take away from this workshop is that repeated measures and longitudinal data analyses need not be intimidating. Put simply, this type of analysis could be thought of as any other regression model, except that we are accounting for non-independence by partitioning the error into an error attributed to the subject and random error. There are many great references out there on modeling longitudinal and repeated measures data. Readers should be encouraged to seek out more information from any of the references listed below and certainly from SAS documentation at https://support.sas.com

REFERENCES

SAS Topics, Repeated Measures Analysis: IDRE, Institute for Digital Research and Education, UCLA, Statistical Consulting Group. http://www.ats.ucla.edu/stat/sas/topics/repeated_measures.htm

Weiss, R. Modeling Longitudinal Data. Springer-Verlag New York, 2005.

Weiss, R. Faculty web site for Rob Weiss which links to homework and labs from his longitudinal and repeated measures courses and data sets from his book on longitudinal analyses: https://faculty.biostat.ucla.edu/robweiss/

CONTACT INFORMATION

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

Leanne Goldstein
City of Hope, Department of Information Sciences, Division of Biostatistics
1500 East Duarte Road
Duarte, CA 91010
Tel: (626) 256-4673
Email: lgoldstein@coh.org

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.