Modeling Longitudinal and Multilevel Data in SAS

Niloofar Ramezani, University of Northern Colorado, Greeley, Colorado

Notice: This is a working draft and more will be added to it later.

ABSTRACT

Correlated data are extensively used across disciplines when modeling data with any type of correlation that may exist among observations due to clustering or repeated measurements.

When modeling clustered data, Hierarchical linear modeling (HLM) is a popular multilevel modeling technique which is widely used in different fields such as education and health studies (Gibson & Olejnik, 2003). A typical example of multilevel data involves students nested within classrooms that behave similarly due to shared situational factors. Ignoring their correlation may result in underestimated standard errors and inflated type-I error (Raudenbush & Bryk, 2002).

When modeling longitudinal data, many studies have been conducted on continuous outcomes; however, fewer studies on discrete responses over time have been completed. These studies require models within conditional, transitional and marginal models (Firzmaurice et al., 2009). Examples of such models which enable researchers to account for the autocorrelation among repeated observations include Generalized Linear Mixed Model (GLMM), Generalized Estimating Equations (GEE), Alternating Logistic Regression (ALR) and Fixed Effects with Conditional Logit Analysis.

This study explores the aforementioned methods as well as several other correlated modeling options for longitudinal and hierarchical data within SAS 9.4 using real data sets. These procedures include PROC GLIMMIX, PROC GENMOD, PROC NLMIXED, PROC GEE, PROC PHREG and PROC MIXED.

Key words: Longitudinal, Hierarchical, Correlated, Discrete Response, GEE

INTRODUCTION

In the presence of multiple time point measurements for subjects of a study and the interest of measuring changes of response variables over time, longitudinal data are formed with the main characteristics of dependence among repeated measurements per subject (Liang & Zeger, 1986). This correlation among measures for each subject makes it inappropriate to keep using the regular modeling options which are used for modeling cross-sectional data with only one measurement per subject. The complexity which is introduced to the study because of the correlated nature of the longitudinal data requires more complex models that enable researchers to take into consideration all aspects of such models. Therefore, more advanced models are required to account for the dependence between multiple outcome values observed for each subject at different time points. Not considering the correlation among observations and continuing using some flawed models for analyzing such data will result in unreliable conclusions.

Throughout years, different models have been proposed to model such data. In the late 1970s, the most common way of modifying ANOVA for longitudinal studies was repeated measures ANOVA which simply models the change of measurements over time through partitioning of the total variation. It was in the early 1980s that Laird and Ware (1982) proposed the linear mixed-effect models for longitudinal data. This class of mixed-effect models has more efficient likelihood based estimations of the model parameters compared to repeated measures ANOVA which makes it a desirable modeling technique (Fitzmaurice, Davidian, Verbeke & Molenberghs, 2009).

Although it has been about a century since the beginning of the development of longitudinal methods for continuous responses, many of the advances in analysis techniques for longitudinal discrete responses have been limited to the recent 30 to 35 years. When the response variables are discrete within a longitudinal study, linear models are no longer appropriate (Firzmaurice et. Al., 2009). To solve this problem, some approximations of GLM have been developed for longitudinal data (Nelder and Wedderburn, 1972). These extensions can be categorized into three main categories: (i) conditional models also known as random-effects or subject-specific models, (ii) transition models, and (iii) marginal or population averaged models (Firzmaurice et al., 2009).

The main methods used in this paper are conditional and marginal models. Conditional models mainly rely on the addition of random effects to the model to account for the correlation among repeated measures. Marginal models directly incorporate the within-subject association among the repeated observations of the longitudinal data into the marginal response distribution. The principal distinction between conditional and marginal models has often been asserted to depend on whether the regression coefficients are to describe an individual's response or the marginal response to changing covariates, that is, one that does not attempt to control for unobserved subjects' random effects (Lee & Nelder, 2004).

GENERALIZED LINEAR MIXED MODEL (GLMM)

One of the conditional models to use when dealing with the correlated data is the GLMM which is a particular type of mixed-effect models. These models contain fixed effects as well as random effects that usually have a normal distribution. More details about it can be found in Agresti (2007) and detailed example SAS codes can be found in Ramezani (2016).

The GLMM model can be written as

$$\eta = X\beta + Z\gamma,$$

where link function is $g(.) = log_e(\frac{p}{1-p})$ and $g(E(y)) = \eta$.

Assuming there exist a longitudinal dataset called Data with a binary dependent variable called DV and three categorical independent variables and one continuous independent variable respectively called IV1, IV2, IV3, and IV4, GLIMMIX and GENMOD procedures in SAS 9.4 can be used to fit a GLMM to this dataset as below. The call to PROC GLIMMIX is displayed:

```
PROC GLIMMIX DATA=Data;
        CLASS IV1 IV2 IV3 ID_CODE;
        MODEL DV = IV1 IV2 IV3 IV4 / DIST=BIN LINK=LOGIT SOLUTION;
        RANDOM INTERCEPT / SUBJECT=ID_CODE;
RUN;
```

Notice all the categorical variables need to be listed within the CLASS statement. Because the dependent variable is binary, the DV will have a binomial distribution and a logistic type of model is appropriate to fit the model. So, within this procedure, options DIST=BIN and LINK=LOGIT are provided to specify a logistic regression model using a generalized linear model link function. Adding the option SUBJECT=ID_CODE to the code will help SAS to recognize the repeated measures that exist for every ID_CODE, hence taking into consideration the dependence among the multiple measures per subject.

For the sample code mentioned above, only intercept is being specified as random but random slope can be used for this model as well by simply adding the name of the respective independent variables in front of the RANDOM statement within PROC GLIMMIX.

If some non-convergence issues happen while fitting the mixed-effect models depending on the data sets which are being used, PROC NLMIXED can be used that has more flexibility under these circumstances.

GENERALIZED ESTIMATING EQUATIONS (GEE)

Liang and Zeger (1986) developed the GEE which is a marginal approach that estimates the regression coefficients without completely specifying the response distribution. In this approach, a 'working' correlation structure for the correlation between a subject's repeated measurements is proposed. Notice that the GEE estimation technique is not a maximum likelihood method.

When modeling discrete response variables, GEE can be used to model correlated data with binary responses. GEE is also appropriate for modeling correlated responses with more than two possible outcomes as well. This technique is a common choice for marginal modeling of ordinal responses for correlated data if the main interest is estimating the regression parameters rather than variance-covariance structure of the longitudinal data. This is because within GEE, the covariance structure is considered as nuisance. The desirable characteristic of a GEE models is that the estimators of the regression coefficients and their standard errors based on GEE are consistent even if the covariance structure for the data is misspecified. GEE is also advantageous because it allows missing values within a subject without losing all the information from that subject. This will result in a higher power of the study which is what every researcher or practitioner is looking for.

The most commonly used within-subject correlation matrices can be categorized into four categories. Independent which suggests no correlation among the repeated observations exists, exchangeable which is used when correlation between any two responses of each subject is the same, autoregressive which assumes the interval length is the same between any two observations, and finally unstructured which suggests there is some sort of correlation between any two responses but the type of correlation is unknown.

GENMOD procedure can be used to fit GEE models for both binary and categorical correlated outcomes. Considering the dataset and variables introduced above, the procedure may be performed as below:

```
PROC GENMOD DATA= Data DESCENDING;
        CLASS IV1 IV2 IV3 ID_CODE;
        MODEL DV = IV1 IV2 IV3 IV4/ DIST=BIN CORRB;
        REPEATED SUBJECT=ID_CODE / CORR=UN;
RUN;
```

The REPEATED statement indicates the use of GEE approach to account for the correlation among repeated observations and CORR=UN specifies an unstructured within-time correlation matrix which can be replaced by other structures.

If the repeated observations were fixed for every subject in this study at the same number of time points, the option CORRW could be added. This option would be used to specify the correlation that exists between the time point measurements in the output. When working with unbalanced data sets, meaning that not every subject has the same number of repeated observations, adding this option is not necessary.

As described above, GEE is also appropriate when modeling longitudinal data with categorical outcomes. When PROC GENMOD was used above for binary outcome, DIST=BIN option was used. To fit the GEE model to categorical outcome variables, the DIST=MULT option must be used within the MODEL statement to request ordinal multinomial logistic modeling option. Assuming that for this example, DV represents a categorical response variable with more than two categories, PROC GENMOD may be performed as below:

```
PROC GENMOD DATA=Data RORDER=data DESCENDING;
        CLASS DV (REF="1") IV1 IV2 IV3 subject_ID;
        MODEL DV= IV1 IV2 IV3 IV4 / DIST=MULTINOMIAL LINK=CUMLOGIT;
        REPEATED SUBJECT=subject_ID / CORR=UN;
RUN;
```

The REF= option in the CLASS statement determines the reference level for EFFECT. This can be used both for categorical dependent and independent variables. The CUMULATIVE link specified within the MODEL statement is referring to the cumulative logit which can be used within these models. More details about different logit functions for modeling categorical response variables along with examples can be found in Ramezani (2016).

PROC GEE is available for modeling ordinal multinomial responses beginning in SAS 9.4 TS1M3. One can use the TYPE= option in the REPEATED statement to specify the correlation structure among the repeated measurements within a subject and fit a GEE to the data as below:

```
PROC GEE DATA= Data DESCENDING;
        CLASS DV (REF="1") IV1 IV2 IV3 subject_ID visit;
        MODEL DV= IV1 IV2 IV3 IV4 visit/ DIST=MULTINOMIAL;
        REPEATED SUBJECT=subject_ID / WITHIN=visit;
RUN;
```

Variable “visit” is being added to this procedure only to be used in the WITHIN option to specify the order of the measurements being recorded in multiple visits or appearances of subjects in the study. Each distinct level of the within-subject-effect defines a different response measures from the same subject. If the data are entered in proper order within each subject, specifying this option is not necessary anymore. If some measurements do not appear in the data for some subjects, this option takes care of this issue by properly ordering the existing measurements and treating the omitted measures as missing values. If the WITHIN= option is not specified for the standard GEE method, missing values are assumed to be the last values and the remaining observations are then ordered in the sequence in which they are provided in the original input data set.

GENERALIZED METHOD OF MOMENTS: A MARGINAL MODEL

GMM, which is also a marginal model, was first introduced in the econometrics literature by Lars Hansen (1982). From then, it has been developed and widely used by taking advantage of numerous statistical inference techniques. Unlike maximum likelihood estimation, GMM does not require complete knowledge of the distribution of the data. Only specified moments derived from an underlying model are what a GMM estimator needs to estimate the model's parameters. This method, under some circumstances, is even superior to maximum likelihood estimator, which is one of the best available estimators for the classical statistics paradigm since the 20th century, as it does not require the distribution of the data to be completely and correctly specified (Hall, 2005). This type of model can be considered a semi-parametric method since the parameter of interest is finite-dimensional and at the same time the full shape of the distributional functions of data may not be known. Within GMM, a certain number of moment conditions, which are functions of the data and the model parameters, need to be specified. These moment conditions have the expected value of zero at the true values of the parameters.

According to Hansen (2007), GMM estimation procedure begins with a vector of population moment conditions taking the form below

$$E[f(\boldsymbol{x}_{it}, \boldsymbol{\beta}_0)] = 0,$$

for all t where $\boldsymbol{\beta}_0$ is an unknown vector of parameters, $\boldsymbol{x}_{it}$ is a vector of random variables, and $f(.)$ is a vector of functions. The GMM estimator is the value of $\boldsymbol{\beta}$ which minimizes a quadratic form in weighting matrix,$\boldsymbol{W}$, and the sample moment $n^{-1}\sum_{i=1}^{n} f(\boldsymbol{x}_{it}, \boldsymbol{\beta})$. This quadratic form is shown as below:

$$Q(\boldsymbol{\beta}) = \{n^{-1}\sum_{i=1}^{n} f(\boldsymbol{x}_{it}, \boldsymbol{\beta})\}'\boldsymbol{W}\{n^{-1}\sum_{i=1}^{n} f(\boldsymbol{x}_{it}, \boldsymbol{\beta})\}.$$

Finally, the GMM estimator of $\boldsymbol{\beta}_0$ is

$$\widehat{\boldsymbol{\beta}} = arg \min_{\boldsymbol{\beta} \in \mathbb{P}} Q(\boldsymbol{\beta}),$$

where $arg\, min$ specifies the value of the argument $\boldsymbol{\beta}$ which minimizes the function in front of it. There are multiple ways of estimating parameters in GMM which details about them can be found in Hall (2005).

PROC MODEL in SAS as well as multiple existing macros can be used to fit a GMM. Iterated generalized method of moments (ITGMM) is one of the ways to fit a GMM to the data. Within ITGMM, the variance matrix for GMM estimation is continuously reestimated at each iteration and this iterative procedure stops when the variance matrix for the equation errors change less than the CONVERGE= value. ITGMM is selected by the ITGMM option within the FIT statement. Simulated Method of Moments (SMM) is another way of fitting GMM via using simulation techniques in model estimation. This method is appropriate for situations in which one deals with the transformation of a latent model into an observable model, random coefficients, missing data, and some other complex scenarios due to the complications involved in the data structure. When the moment conditions are not readily available in closed forms but can be approximated via simulation, simulated generalized method of moments (SGMM) can be used to fit a GMM to the data. Using the example from chapter 19 of the SAS/ETS ® 13.2 User's Guide (2014), suppose one is interested in using GMM for estimating the parameters of this model

$$y = a + bx + u,$$

Where u follows a normal distribution with the mean of zero and variance of s^2. Specifying the first two moments of it as $E(y) = a + bx$ and $E(y^2) = (a + bx)^2 + s^2$ will result in _eq.m1 = y-(a+b*x)_ and _eq.m2 = y*y - (a+b*x)**2 - s*s_ to be used within the PROC MODEL statement as the moment equations. This model can be estimated by using GMM with following statements:

```
PROC MODEL DATA= Data;
        PARMS a b s;
        INSTRUMENT x;
        eq.m1 = y-(a+b*x);
        eq.m2 = y*y - (a+b*x)**2 - s*s;
        BOUND s > 0;
        FIT m1 m2 / gmm;
RUN;
```

The gmm option specified within the FIT statement results in the use of the GMM for parameter estimation. If the closed form for the moment conditions is not available, the moment conditions can be simulated by generating simulated samples based on the parameters. Using SGMM as below will result in the desired estimated parameters:

```
PROC MODEL DATA= Data;
        PARMS a b s;
        INSTRUMENT x;
        ysim = (a+b*x) + s * rannor( 8003);
        y = ysim;
        eq.ysq = y*y - ysim*ysim;
        FIT y ysq/ gmm ndraw;
        BOUND s > 0;
RUN;
```

Unfortunately, the existing procedures are not very straightforward and developing easy-to-use procedures in SAS to fit GMM models can encourage applied researchers to use these techniques which due to their complexities are not widely used in different fields. It is important to provide simple programming tools for such models so they could be easily adopted when needed as their higher efficiency compared to other models has been proven when modeling longitudinal data especially in the presence of time dependent covariates (Lai & Small, 2007).

ALTERNATING LOGISTIC REGRESSION (ALR)

ALR is another marginal model that models the association among repeated measures of the response variable with odds ratios rather than correlations which is what GEE uses. Similar to GEE, the ALR model provides estimates of the marginal model parameters; however, it does not restrict the type of correlation among the repeated measurements as the GEE method does. This use of odds ratios will result in parameter estimates of the model on the log odds ratios among the measurements.

PROC GEE can also be used for fitting an ALR to a data set. When trying to fit the ALR method, using the option LOGOR= is required and TYPE= should not be specified anymore. PROC GEE uses TYPE=IND by default within the GEE procedure trying to exclude the correlation restriction of the GEE model which is necessary for the ALR method. The ALR model can be fitted as below:

```
PROC GEE DATA= Data DESCENDING;
        CLASS DV (REF="1") IV1 IV2 IV3 subject_ID visit;
        MODEL DV= IV1 IV2 IV3 IV4 visit/ DIST=MULTINOMIAL;
        REPEATED SUBJECT=subject_ID / WITHIN=visit LOGOR=EXCH;
RUN;
```

Specifying LOGOR=EXCH in the REPEATED statement selects the fully exchangeable model for the log odds ratio of an ALR model. The same specifications may be made within PROC GENMOD to fit the same ALR model.

As mentioned in detail in Ramezani (2016), in the presence of missing data, PROC GEE can also be used to implement the weighted GEE method when missing responses depend on previous responses which is not available in PROC GENMOD. A GEE model, estimated by residual pseudo-likelihood, can be fitted using the GLIMMIX procedure by specifying the EMPIRICAL option in the PROC GLIMMIX statement. Furthermore, specifying the RANDOM _RESIDUAL_ statement with the subject variable in the SUBJECT= option is required.

FIXED EFFECTS WITH CONDITIONAL LOGIT ANALYSIS

Fixed effects with conditional logit analysis is a conditional modeling technique for correlated observations which treats each measurement of each subject as a separate observation. This subject-specific model is more appropriate for binary correlated outcomes rather than multinomial correlated dependent variables. But this model has some advantages which make it beneficial to use it for modeling longitudinal ordinal outcomes by using the cumulative logit which in fact dichotomizes multiple categories of the response. GEE does not correct for the bias resulting from omitted explanatory variables of the cluster level; therefore, by adding an intercept, α_i, to the model, one can statistically control for all stable characteristics of subjects of the study (Allison, 2012). This term implements a positive correlation among the observed outcome. The general model will be as below

$$\log\left(\frac{p_{it}}{1-p_{it}}\right)=\alpha_i+\beta x_{it},$$

where the logit function represents the logit of the probability of having a outcome of interest for subject i, p_{it}, at time point t and α_i represents all differences among individuals that are stable over time. If this term is treated as a random effect with any distribution such as normal, this model will be a random-effects or mixed-effect model and can be modeled using GLIMMIX macro or PROC NLMIXED.

The fixed effects with conditional logit analysis may be fitted in SAS through a PHREG procedure due to the similarity between the likelihood function used in this model and the likelihood function for stratified Cox regression analysis. The sample code may be written as below:

```
PROC PHREG DATA= Data;
        MODEL DV= IV1 IV2 IV3 IV4 / TIES=DISCRETE;
        STRATA subject_ID;
RUN;
```

TIES=DISCRETE is added because of the different categories of the response that subjects fall into each time. In case of them falling into the same categories at each time point, this option would be unnecessary.

MULTILEVEL MODELS OR HIERARCHICAL LINEAR MODELS

Multilevel data structures are common in studies with some kind of clustering among subjects of the study. A typical example of multilevel data involves students nested within classrooms or patients clustered within hospitals. Students belonging to the same class or patients within the same hospital may respond in similar ways to an outcome measure due to shared situational factors. This will result in having correlated data, hence violating the independence assumption of observations. Hierarchical linear modeling (HLM) is a popular multilevel modeling technique which accounts for the variability nested with clusters by allowing for simultaneously inclusion of parameters across levels (Gibson & Olejnik, 2003). More details about these models can be found in Raudenbush and Bryk (2002).

The most basic HLM is a one-way random effect ANOVA in which intercept variation is estimated across some grouping factor. Specification of the model can take two forms. The combined equation is the one which is being shown here for the basic model which simply is a regression formula but with the addition of a random effect term to represent the cluster level deviation from the outcome mean. This model can be shown as below:

$$Y_{ij}=\gamma_{00}+u_{0j}+r_{ij},$$

where Y_{ij} represents an outcome for unit i within cluster j, γ_{00} is the grand mean for Y_{ij}, u_{0j} is a random effect representing the cluster deviation from the grand mean and finally r_{ij} is another random effect term which represents the individual deviation from the grand mean.

A more complex HLM which is referred to as a random-coefficients model is an extension of the basic model which was explained above. This model has predictors and random slope terms at different levels of the multilevel model. This morel can be specified as below in separate equations:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(x_{ij}) + \ r_{ij},$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(W_j) + u_{0j},$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(W_j) + u_{1j},$$

where β_{0j} is the randomly varying intercept term. β_{0j} can be modeled as a linear combination of the grand mean of the response variable, γ_{00}, a fixed effect, γ_{01}, and a residual term, u_{0j}. Any number of predictors, specified as γ_{01} through γ_{0n}, can be added to this equation to model variation in the intercept across clusters. β_{1j} which is the randomly varying slope term can be modeled as the linear combination of a grand mean, γ_{10}, representing the average relationship between a level one predictor and the dependent outcome, a fixed effect predictor γ_{11} and a residual term u_{1j}. Again, any number of available predictors, denoted γ_{11} through γ_{1n}, can be included in the model.

Here is the combined equation for the random-coefficients model:

$$Y_{ij} = \gamma_{00} + \ \gamma_{01}(W_j) + \gamma_{10}(x_{ij}) + \gamma_{11}(W_j * x_{ij}) + u_{0j} + u_{1j}(x_{ij}) + r_{ij}$$

Suppose we have a multilevel data called Data. Referring to the school example mentioned earlier with students clustered within school, the outcome variable, DV, can be a student-level variable such as GPA. Variable IV is a student-level variable such as social-economic-status of a student. Variable MEANIV is a school-level variable which is the group mean of the student level social-economic-status. Both IV and MEANIV are centered at the grand mean (having the means of 0). There are different types of HLM models which can be built but the one which is of interest here is the random-coefficients model based on Including the effects of the student-level predictors in the model and trying to predict the DV from the centered student-level IV.

Within SAS, PROC MIXED can be used to fit an HLM. But because here we have decided to fit a random-coefficients model, first the centered IV needs to be defined which is IVc=IV-MEANIV. Using this variable in the model, the SAS code can be written as below:

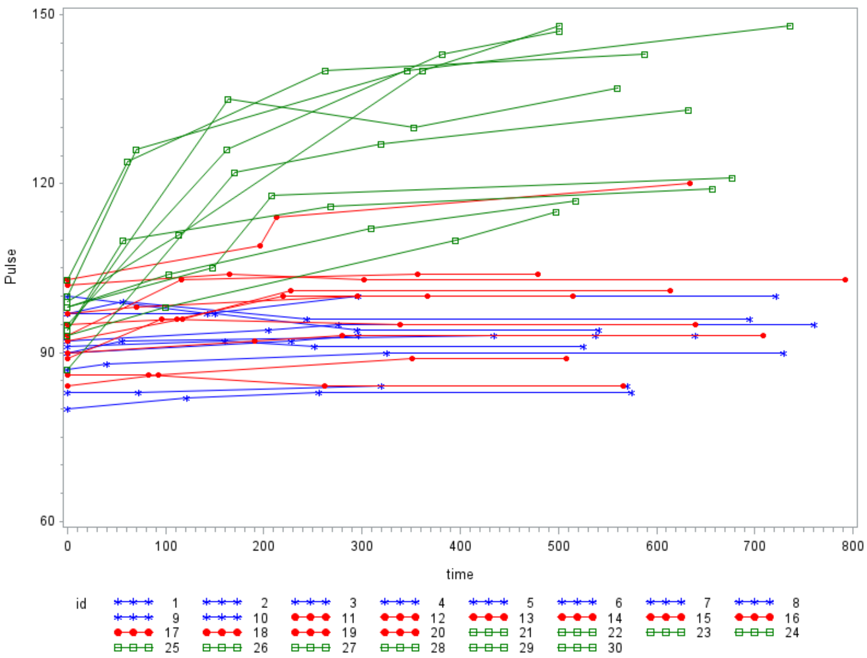
```
PROC MIXED DATA= Data;
        CLASS SCHOOL;
        MODEL DV= IVc / SOLUTION  DDFM=BW;
        RANDOM intercept IVc / SUBJECT =SCHOOL TYPE =UN GCORR;
RUN;
```

Within PROC MIXED, option TYPE =UN in the RANDOM statement allows estimating parameters for the variance of intercept, the variance of slopes for IVc and the covariance between them from the data. Option GCORR is added to display the correlation matrix called G matrix which corresponds to the estimated variance-covariance matrix. SOLUTION in the MODEL statement gives the parameter estimates for the fixed effect. Option DDFM = BW is added in the same statement to request SAS to use between and within method for computing the denominator degrees of freedom for the tests of fixed effects. This option is especially useful in the presence of lots of random effects in the model and when the design is severely unbalanced. The default should be used for the tests performed for balanced split-plot designs and can also be used for moderately unbalanced designs.

There are some other types of multilevel models which are called growth curve models. Within such models, longitudinal observations are nested within each individual or subject. They are useful in a way that as multilevel models, they estimate changes within subjects over time and at the same time, they can summarize the differences that exist between individuals (Little, 2013). Modeling such models can be done using PROC MIXED and time can be used as a predictor. Like GEE, within MIXED procedure, different types of correlations among repeated observations can be specified such as compound symmetry (CS), unstructured (US) and autoregressive (AR(1)).

DATA ANALYSIS

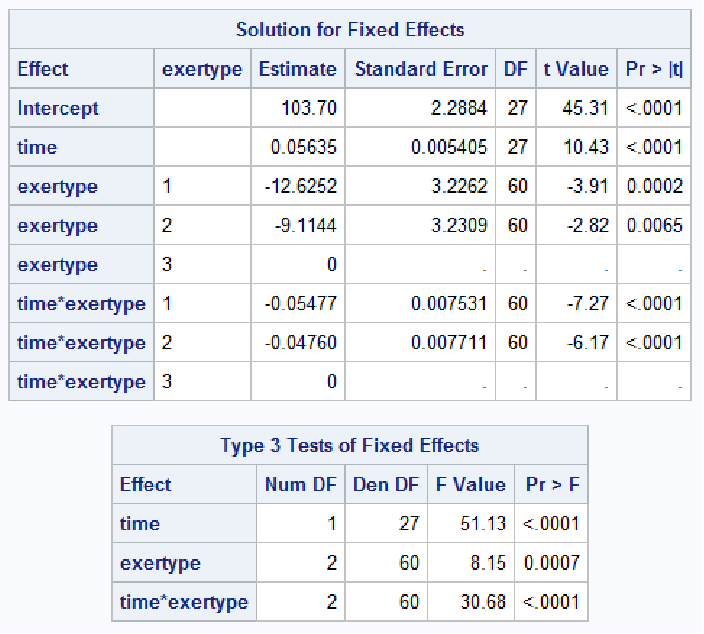
Looking at the pulse rate of some patients over time, their diet type and exercise type have been recorded. Here is the general trend of each patient over time:



Fitting the correlated model to the continuous outcome as below

```
PROC GENMOD DATA=study2 RORDER=data DESCENDING;
CLASS id exertype;
MODEL pulse = time exertype time*exertype;
        REPEATED SUBJECT=id /CORR=UN;
RUN;
```

The output will be presented as below



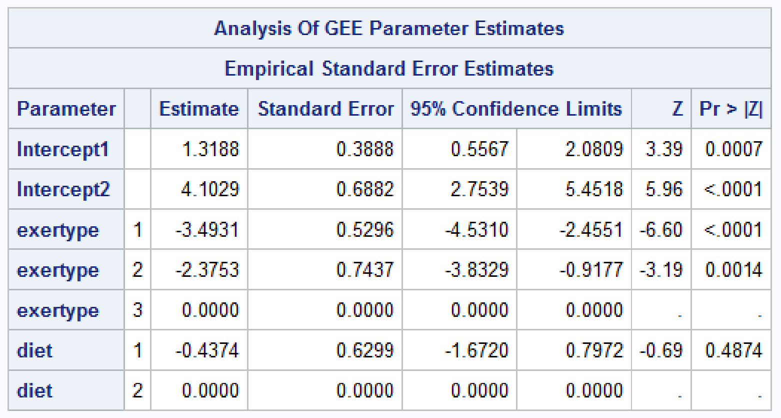
Solution for Fixed Effects						
Effect	exertype	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		103.70	2.2884	27	45.31	<.0001
time		0.05635	0.005405	27	10.43	<.0001
exertype	1	-12.6252	3.2262	60	-3.91	0.0002
exertype	2	-9.1144	3.2309	60	-2.82	0.0065
exertype	3	0	.	.	.	.
time*exertype	1	-0.05477	0.007531	60	-7.27	<.0001
time*exertype	2	-0.04760	0.007711	60	-6.17	<.0001
time*exertype	3	0	.	.	.	.

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
time	1	27	51.13	<.0001
exertype	2	60	8.15	0.0007
time*exertype	2	60	30.68	<.0001

Looking at the categorical outcome, the model can be fitted as below:

```
/*Creating the categorical variable for pulse*/
DATA study2;
SET study2;
IF pulse<=90 THEN pulse2=1;
IF (pulse>=91 and pulse<=100) THEN pulse2=2;
IF pulse>101 THEN pulse2=3;
RUN;
PROC GENMOD DATA=study2 RORDER=data DESCENDING;
CLASS pulse2 id exertype diet;
MODEL pulse2 =  exertype diet / DIST=MULTINOMIAL LINK=CUMLOGIT;
        REPEATED SUBJECT=id / CORR=independent;
RUN;
```

The output will be presented as below for a GEE model:



Analysis Of GEE Parameter Estimates

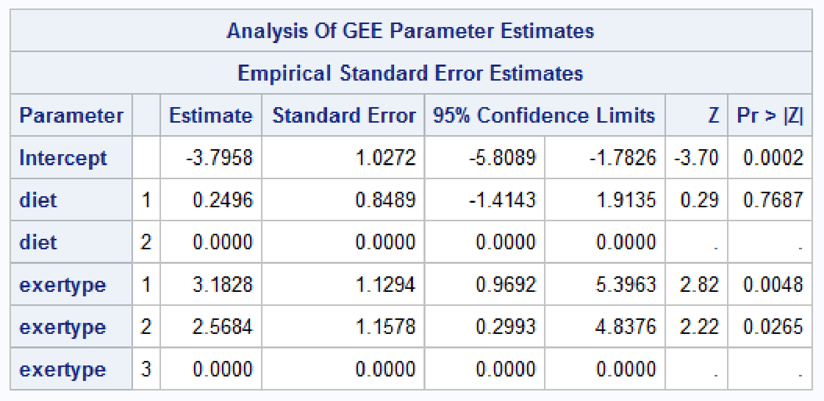
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept1		1.3188	0.3888	0.5567	2.0809	3.39	0.0007
Intercept2		4.1029	0.6882	2.7539	5.4518	5.96	<.0001
exertype	1	-3.4931	0.5296	-4.5310	-2.4551	-6.60	<.0001
exertype	2	-2.3753	0.7437	-3.8329	-0.9177	-3.19	0.0014
exertype	3	0.0000	0.0000	0.0000	0.0000	.	.
diet	1	-0.4374	0.6299	-1.6720	0.7972	-0.69	0.4874
diet	2	0.0000	0.0000	0.0000	0.0000	.	.

Having a binary outcome, the code can be written as below:

```
/*Creating the binary variable for pulse*/
DATA study2;
SET study2;
IF pulse<=90 THEN pulse3=1;
IF pulse>91 THEN pulse3=2;
RUN;
PROC GENMOD DATA=study2;
CLASS id exertype diet;
MODEL pulse3 = diet exertype / DIST=BIN LINK=LOGIT;
        REPEATED SUBJECT=id;
RUN;
```

Another GEE model is fitted to the data and the output is as below:



Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-3.7958	1.0272	-5.8089	-1.7826	-3.70	0.0002
diet	1	0.2496	0.8489	-1.4143	1.9135	0.29	0.7687
diet	2	0.0000	0.0000	0.0000	0.0000	.	.
exertype	1	3.1828	1.1294	0.9692	5.3963	2.82	0.0048
exertype	2	2.5684	1.1578	0.2993	4.8376	2.22	0.0265
exertype	3	0.0000	0.0000	0.0000	0.0000	.	.

CONCLUSION

Different options for modeling longitudinal binary, multinomial and ordinal responses were discussed above. Procedures developed within SAS such as PROC GENMOD, PROC GLIMMIX, PROC NLMIXED, PROC PHREQ and PROC GEE can appropriately model dichotomous correlated outcomes. If the outcomes are not longitudinal, PROC LOGISTIC can easily handle different types of logistic models for modeling cross-sectional binary, multinomial or ordinal response variables.

At the end, some multilevel models were introduced as there exists correlation among observations for clustered data as well which needs to be considered within different modeling options. These models are HLM and growth curve models which can be modeled using PROC MIXED as explained above. The important point which needs to be considered by applied researchers is the necessity of accounting for the correlation that exists among observations when modeling longitudinal or multilevel data. Taking their correlation into consideration is important because the existence of repeated measurements results in the violation of the independence of observations assumption; therefore, the regular models cannot appropriately model longitudinal data. Using appropriate models which were discussed above will result in more informative and powerful models.

Author is currently working on extending this study in different directions such as longitudinal growth-curve modeling using HLM. It is of interest as longitudinal HLM offers an opportunity to have each subject serve as his or her own control within the study to eliminate between-individual third-variable confounds.

REFERENCES

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York: Wiley.

Allison, P. D. (2012). *Logistic regression using SAS: Theory and application*. SAS Institute.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (Eds.). (2009). Longitudinal data analysis. CRC Press.

Gibson, N. M., & Olejnik, S. (2003). Treatment of missing data at the second level of hierarchical linear models. Educational and Psychological Measurement, 63(2), 204-238.

Lee, Y., & Nelder, J. A. (2004). Conditional and marginal models: another view. Statistical Science, 19(2), 219-238.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. Biometrics, 963-974.

Liang, K. Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13-22.

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

Ramezani, N. (2016). Analyzing non-normal binomial and categorical response variables under varying data conditions. In proceedings of the SAS Global Forum Conference. Cary, NC: SAS Institute Inc.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (Vol. 1). Sage.