Paper 1251-2017

Analyzing Correlated Data in SAS®

Niloofar Ramezani, University of Northern Colorado

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 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 (Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2009). Examples of such models which enable researchers to account for the autocorrelation among repeated observations include Generalized Linear Mixed Model, Generalized Estimating Equations, Alternating Logistic Regression 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. These procedures include PROC GLIMMIX, PROC GENMOD, PROC GEE, PROC PHREG, PROC MODEL and PROC MIXED.

INTRODUCTION

In the presence of multiple time points for subjects of a study and the interest of modeling patterns of change over time, longitudinal data are formed. The main characteristic of such correlated data is the dependence that exists among repeated measurements per subject (Liang & Zeger, 1986). This correlation among measures for each subject results in a complexity due to the violation of the independence assumption that should be met for cross-sectional models (Bena & McIntyre, 2008). For example, when modeling students' academic performance over multiple semesters there are multiple measurements for each student. The academic performance of each student may vary through a period of time, but these observations are correlated within each student making cross-sectional data models inappropriate.

Many models have been developed for cross-sectional data where a single observation for each subject is available for only a single time point in the study, but more studies regarding modeling of longitudinal data in the presence of varying types of responses need to be conducted regardless of their challenges. The main reasons conducting more studies in this area is important are their extensive use in different fields, the desirable information modeling changes over time provides for applied researchers and practitioners, and the opportunities repeated observations provide for the researchers such as increased power of the study and robustness to model selection.

The correlation among observations in a study can also be caused by clustering of subjects within groups due to their similarities. For instance, students nested within classrooms may behave similarly in terms of their academic performance due to shared situational factors such as same peers and the same teachers. Whether the correlation among observations in a study is caused by the longitudinal nature of a study or because of clustering, it is crucial to take the correlation into consideration while modeling this kind of data. If not, there is an analytic cost researchers may pay by inconsistent estimates of precision because of ignoring the existing correlation among subjects of longitudinal data. However, these challenges can be overcome by appropriate models that are specifically designed to capture the correlation among the observations and use them to have a greater power and more appropriate inferences. Some of these methods that can be used in modeling longitudinal and in general correlated data are discussed below.

MODELING LONGITUDINAL DATA

The early development of methods that can handle longitudinal data is traced back to the seminal paper by Harville (1977) and the use of ANOVA paradigm for longitudinal studies. After developing the repeated measures ANOVA technique for analyzing correlated data, the idea of adding random effects to a model with fixed effects and ending up with a mixed-effects model was developed. Later, more advanced conditional, transition, and marginal models were established which some are discussed below.

LINEAR MIXED-EFFECT MODELS

It was in the early 1980s that Laird and Ware (1982) proposed their flexible class of linear mixed-effect models for longitudinal data. The linear mixed-effects model is given by

$$Y_{it} = \boldsymbol{X}'_{it}\boldsymbol{\beta} + \boldsymbol{Z}'_{it}\boldsymbol{\gamma} + e_{it},$$

where Y_{it} is the response variable of the ith subject observed repeatedly at different time points $(i = 1,\ldots,n, t = 1,\ldots,T)$, n is the number of subjects, T is the number of time points, $\boldsymbol{X}_{it}$ is the design or covariance vector of tth measurement measured at time t for subject i for the fixed effects, $\boldsymbol{\beta}$ is the fixed effect parameter vector, $\boldsymbol{Z}_{it}$ is the design vector of the tth measurement measured for subject i for the random effects, $\boldsymbol{\gamma}$ is the random effect parameter vector following a normal distribution, and e_{it} is the random error also following a normal distribution. Having the random effects within these mixed-effect models helps account for the correlation among the measurements per subject at different time points.

One of the SAS procedures that can be used to fit such models is PROC MIXED. Using this procedure, random intercept, random slope, or both can be imposed within the model. The call to a random intercept model is displayed:

```
PROC MIXED DATA=Data;
    CLASS ID;
    MODEL DV = time / S CORRB DDFM=SATTERTHWAITE;
    Random Int / TYPE=VC  SUB=ID;
RUN;
```

Where ID is the ID for subjects that have multiple records in the dataset called Data depending on the number of measurements per subject, DV is the response variable, and time is the multiple time points considered in the study. Notice that only intercept is specified within the RANDOM statement but adding a random slope to this model is as easy as adding it to the random statement as below:

```
PROC MIXED DATA=Data;
    CLASS ID;
    MODEL DV = time / S CORRB;
    RANDOM Int time/ TYPE=VC SUB=ID;
RUN;
```

In order to allow for the covariation in parameters, all that needs to be done is using TYPE=unr. If one is interested in fitting a random-slopes model with a non-linear time, for example quadratic time, PROC MIXED can be used again similar to the way it was used above and the quadratic term can be added by using time|time in the MODLE statement. The model is displayed as below:

```
PROC MIXED DATA=Data;
    CLASS ID;
    MODEL DV = time|time / S CORRB DDFM=SATTERTHWAITE;
    RANDOM Int time / TYPE=un  SUB=ID;
RUN;
```

NONLINEAR MODELS

Although many studies have been conducted for about a century regarding analyzing longitudinal continuous responses, many of the advances in the development of methods for analyzing longitudinal discrete responses have been limited to the recent 35 years. When the response variables are discrete, they are no longer normally distributed and therefore linear models are no longer appropriate.

Development of generalized linear model (GLM) and its approximations has solved the issue of modeling non-normal response variables and so GLM has been used for modeling discrete responses. However, the addition of a non-linear transformation of the mean under the assumption that it is a linear function of the covariates within GLM can introduce some issues in the regression coefficients of longitudinal data. This problem has been solved by extending the GLMs to handle longitudinal observations in a number of different ways. Such extended models can be categorized into three main categories of (i) conditional models also known as random-effects or subject-specific models, (ii) transition models, and (iii) marginal or population averaged models (Fitzmaurice et al., 2009).

CONDITIONAL, TRANSITION, AND MARGINAL MODELS

Conditional models are appropriate when examining individual level data is of interest. For example, consider modeling the academic achievement of students clustered into majors within a single school or the academic success of students over time. One will be interested in conditional models if the interpretation of the results seeks to explain what factors impact academic success or achievement of students and their individual trend (Zorn, 2001). Conditional or random effects models allow adding a random term to the model to capture the correlation among the observations. A linear version of conditional models can be specified as

$$Y_{it} = \boldsymbol{X'}_{it}\boldsymbol{\beta} + \nu_i + e_{it},$$

where both the random effect, ν_i, and error term, e_{it}**,** follow a normal distribution with different moments.

Generalized Linear Mixed Model: A Conditional Model

One example of conditional models is Generalized Linear Mixed Model (GLMM) which is an extension of GLM that includes a random effect, and hence can be applied to longitudinal and correlated data. These models contain fixed effects as well as random effects that usually have a normal distribution.

Within GLMM, rather than modeling the responses directly, some link function is often applied, such as a log or logit link. Let the linear predictor, $\boldsymbol{\eta}$, be the combination of the fixed and random effects excluding the residuals specified as below

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma},$$

where $\boldsymbol{\beta}$ is the vector of fixed effects, $\boldsymbol{X}$ is the fixed effects design matrix, $\boldsymbol{\gamma}$ is the vector of random effects, and $\boldsymbol{Z}$ is the random effects design matrix. The link function, $g(.)$, relates the outcome vector, $\boldsymbol{Y}$, to the linear predictor , $\boldsymbol{\eta}$. One of the most common link functions is the logit link in which $g(.) = log_e(\frac{p}{1-p})$ and $g(E(\boldsymbol{Y})) = \boldsymbol{\eta}$. The default optimization technique within such methods is the Quasi-Newton method. More details about these models can be found in Agresti (2007) and detailed example SAS codes can be found in Ramezani (2016).

Multiple procedures within SAS such as PROC GLIMMIX, PROC NLMIXED, and PROC GENMOD can be used in SAS 9.4 to fit GLMMs. Suppose there exist a binary longitudinal response called DV, a categorical predictor called IV1, and a continuous predictor called IV2. The call to PROC GLIMMIX in order to fit a GLMM to the aforementioned dataset is displayed:

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

All categorical variables need to be listed within the CLASS statement. Options DIST=BIN and LINK=LOGIT are provided to specify a binomial distribution for the response and a generalized linear model logit link function. Adding the option SUBJECT=ID to the code is the key for SAS to recognize the repeated measures that exist for every ID. Different distributions and link functions may be used within this procedure depending on the nature of the data used for a study.

Fixed Effects with Conditional Logit: A Conditional Model

Another example of conditional models is fixed effects with conditional logit analysis. This model is based on treating every measurement of each subject as a separate observation. One of the main characteristics of such models is adding an intercept, α_i, to the model to statistically control for all stable characteristics of subjects of the study by implementing a positive correlation among the observed outcome (Allison, 2012). 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 an 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.

The fixed effects with conditional logit analysis can be fitted in SAS using a PHREG procedure which is the procedure also used for survival models. Using the same procedure for both models is due to the similarity between the likelihood function used in fixed effects with conditional logit model and the one used in stratified Cox regression analysis. The sample code may be written as below:

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

TIES=DISCRETE is added because of the different categories of the response that subjects may fall into at different time points. ID is used within the STRATA statement to specify that each subject, which has its own unique ID, is a one stratum with repeated measurements of each subject clustered within it that are correlated with each other. This is how the correlation among the observations of each subject is acknowledged, captured, and taken into account within this procedure.

Transition models are also an extension of generalized linear models by modeling the mean and time dependence simultaneously. Transition or Markov model is a specific kind of conditional models which accounts for the correlation between subjects of a longitudinal study by conditioning an outcome on the other outcomes that lets the past values influence the present observations which are of interest.

Marginal models are also the extension of GLM which directly incorporate the within-subject association among the repeated measures into the marginal response distribution. The main distinction between marginal and conditional models has often been asserted to depend on whether the regression coefficients describe an individual’s response or the marginal response to changing covariates (Lee & Nelder, 2004). Marginal models can be written as below in general

$$E(Y_{it}) = \boldsymbol{X'}_{it}\boldsymbol{\beta},$$

where the parameters in the variances of the responses are nuisance parameters with an arbitrary chosen pattern.

These models include no random effect and are population averaged models such as generalized estimating equations (GEE) and generalized method of moments (GMM). According to Hansen (2007), marginal approaches are appropriate when the researcher seeks to make inferences about the population average (Diggle, Liang, & Zeger, 1994). For the example mentioned above about academic achievement and success of students, if the goal of a study is to compare the academic success between clusters or majors, a marginal model would be the appropriate model to fit to the data (Zorn, 2001).

Generalized Estimating Equations: A Marginal Model

GEE is a marginal model developed by Liang and Zeger (1986) to estimates the regression coefficients without completely specifying the response distribution. The use of a ‘working’ correlation structure to describe the correlation between a subject’s repeated measurements is the main characteristic of GEE models. The most commonly used within-subject correlation matrices are ‘independent’ which suggests no correlation among the repeated observations, ‘exchangeable’ which suggests the same correlation between any two responses of each subject, ‘autoregressive’ which assumes the same interval length between any two observations, and ‘unstructured’ which suggests an unknown correlation between any two responses.

When modeling discrete response variables, GEE can be used to model correlated data with binary and multinomial responses. 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. Procedures such as GENMOD and GEE can be used within SAS 9.4 to fit a GEE model to both binary and categorical correlated outcomes as below.

GENMOD procedure can be used to fit GEE models for both binary and categorical correlated outcomes through specifying the distribution of the response of interest using the DIST option within MODEL statement. Considering the dataset and variables introduced above, the procedure can be performed as below:

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

By using the REPEATED statement, the use of GEE approach is imposed and CORR=UN specifies an unstructured within-time correlation matrix which can be replaced by other correlation structures mentioned above. To fit the GEE model to categorical responses with more than two categories using PROC GENMOD, the DIST=MULT option must be used within the MODEL statement to request multinomial logistic modeling option and a LINK option should also be added within the MODEL statement. For more details about different logit functions and example SAS codes for each logit link when modeling categorical outcome variables using ordinal and multinomial logistic models, please see Ramezani (2016).

PROC GEE is also available to fit GEE models to correlated data 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 ID visit;
        MODEL DV= IV1 IV2 visit/ DIST=MULTINOMIAL;
        REPEATED SUBJECT=ID / TYPE =un WITHIN=visit;
RUN;
```

Here the unstructured correlation structure is used but any other type of structure may be imposed within this procedure depending on the nature of the data and the correlation that exists among observations. Variable “visit” is being added to be used in the WITHIN option to specify the order of the measurements in multiple appearances of subjects in the study. If some measurements are missing in the data for some subjects, this option takes care of this issue by properly ordering the measurements and treating the omitted measures as missing values. GLIMMIX procedure can also be used to fit a GEE model, estimated by residual pseudo-likelihood, by specifying the EMPIRICAL option.

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: A Marginal Model

Alternating Logistic Regression (ALR) is another marginal model that models the correlation among repeated measures of the response variable with odds ratios rather than correlations, which is what GEE uses, or moments, which is what GMM uses. The ALR model provides estimates of the marginal model parameters without restricting the type of correlation among the repeated measurements. PROC GENMOD and PROC GEE can be used for performing an ALR. When trying to fit the ALR method within PROC GEE, using the option LOGOR= is required. TYPE=IND is the default for the GEE procedure when fitting ALR in an effort to exclude any correlation restriction. The ALR model can be fitted as below:

```
PROC GEE DATA= Data DESCENDING;
        CLASS DV (REF="1") IV1 ID visit;
        MODEL DV= IV1 IV2 / DIST=MULTINOMIAL;
        REPEATED 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.

MODELING CORRELATED MULTILEVEL DATA

Multilevel data structures should be used when there exists clustering among subjects of the study. A typical example of multilevel data involves students nested within classrooms which may respond in similar ways to an outcome measure due to shared situational factors such as their teacher or classroom environment. The correlation that exists among the students of the same classroom will result in violating the independence assumption of observations. Hierarchical linear modeling (HLM) is a common and effective technique which accounts for the variability nested with clusters (Gibson & Olejnik, 2003). More details about these models can be found in Raudenbush and Bryk (2002).

A one-way random effect model is the basic HLM model in which intercept variation is estimated across some grouping factor is shown below. The added random effect term represents 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 the response, u_{0j} is a random effect representing how much difference exist between each cluster and the grand mean, and r_{ij} is the random effect term representing how much difference exists between each individual and the grand mean.

Random-coefficients HLM model, which is an extension of the basic model, includes predictors and random slope terms at different levels of the multilevel model. This morel can be specified as below:

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

where γ_{00} is the grand mean of the response variable, γ_{01} is a fixed effect, and u_{0j} is a residual term which are all used to model a randomly varying intercept term within such models. Any number of predictors can be added to this equation to model variation in the intercept across clusters. γ_{10} is the average relationship between a level one predictor and the dependent outcome, also referred to as the grand mean, γ_{11} is a fixed effect predictor, and u_{1j} is a residual term which all are used to model the randomly varying slope term. Again, any number of available predictors can be included in this model.

PROC MIXED can be used to fit a multilevel or HLM model in SAS as below:

```
PROC MIXED DATA= Data;
        CLASS CLASSROOM;
        MODEL GPA= IQc / SOLUTION  DDFM=BW;
        RANDOM intercept IQc / SUBJECT = CLASSROOM TYPE =UN GCORR;
RUN;
```

Data is a multilevel data used for this procedure. Referring to the educational example mentioned earlier with students clustered within classrooms, the outcome variable is a student-level achievement variable such as GPA. The student-level independent variable used here is the IQ of a student and IQc is the IQ variable centered at the grand mean.

Option TYPE =UN in the RANDOM statement allows estimating parameters for the variance of intercept, the variance of slopes for the independent variable, and the covariance between them from the data. Option GCORR will display the correlation matrix which corresponds to the estimated variance-covariance matrix. SOLUTION within the MODEL statement gives the parameter estimates for the fixed effect. Option DDFM = BW is also added in the MODEL statement to request SAS to use between and within method for computing the denominator degrees of freedom for the tests of fixed effects.

There exist other types of multilevel models which are called growth curve models to estimate changes within subjects over time as well as the differences among individuals (Little, 2013). PROC MIXED can also be used for fitting such models.

GRAPHICAL PRESENTATION OF LONGITUDINAL DATA

The importance of producing descriptive statistics and graphs to help understanding and interpreting different trends and changes over time in longitudinal studies should not be underestimated. A few helpful tools are introduced below and many more exist which using them is recommended to enable the researcher or data analyst to describe the characteristics of the longitudinal data as well as modeling them using the methods mentioned in previous sections of this paper.

Time plot is one of these helpful tools which can be used to evaluate the patterns and behavior of the data over time. PROC SGPLOT can be used to produce this plot as below where DV is the measured response variable of interest and time includes the information regarding the multiple time points of measurement in the study:

```
PROC SGPLOT DATA=Data;
        SCATTER y=DV x=time;
        TITLE 'Time plot';
RUN;
```

Figure 1 shows an example of a time plot which indicates a growth in the measures of the dependent variable over time. As obviously seen, this plot clarifies different patterns that exist in data over time which is a great descriptive tool for interpreting such patterns.

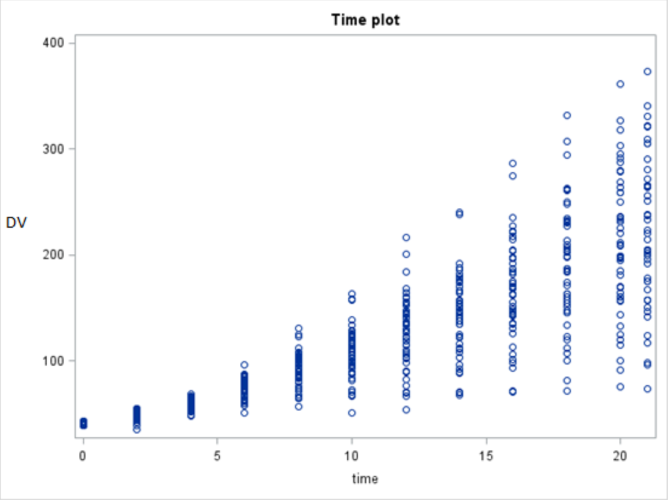

Figure 1. Time Plot

Spaghetti plot is another tool to visualize changes over time per subject or the possible flows through systems over time. PROC SGPLOT can be used to produce this plot as below by using the SERIES statement instead of SCATTER which was used within the SGPLOT procedure when producing time plots. Below is the SAS code for producing a simple spaghetti plot:

```
PROC SGPLOT DATA=Data;
        SERIES y=DV x=time / GROUP=ID;
        TITLE 'Spaghetti plot';
RUN;
```

Where GROUP=ID is provided to look at each subject and show the changes of the responses for each subject over time using one line, also referred to as a noodle. Below, an example of a spaghetti plot is provided that shows the growth of the measured response variable in time per subject.

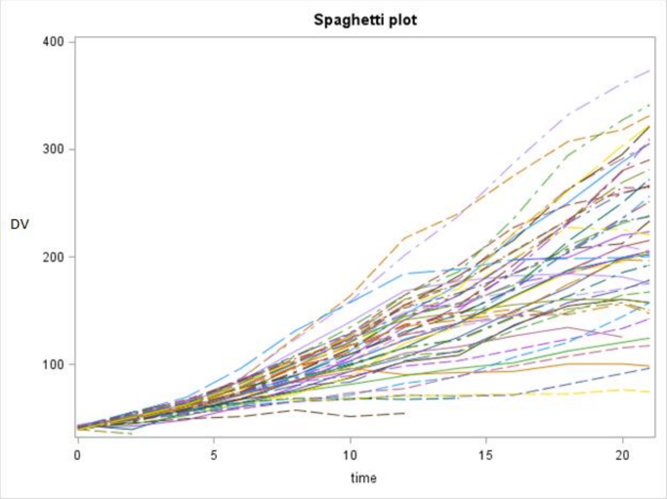

Figure 2. Spaghetti Plot

Interaction plot is another tool to visualize the interactions between two variables which using them within longitudinal data can reveal how variables interact with each other over time. PROC SGPLOT can be used to produce this plot but it requires a PROC SORT and a PROC UNIVARIATE before running the SGPLOT procedure.

Assume there is an independent variable with four different levels; for example different methods of teaching used for different cohorts of students. Suppose we are interested to see how these different methods affect students' achievement over time for this example and in general how different levels of variables might interact with each other over time in terms of the changes they make in the response. The example SAS code below shows how to produce and interaction plot over time to get a separate line for the means of each level of the independent variable against the dependent variable. As mentioned above, there are three steps for creating the interaction plots. First, the data needs to be sorted by the independent variable and time through a SORT procedure. Second, PROC UNIVARIATE needs to be used for finding the means of the responses by the independent variable at different time points as shown below. Finally, PROC SGPOT is used by specifying the GROUP=IV as this time the grouping factor is the independent variable and not the individual subjects as they were in a spaghetti plot. Notice, within PROC SGPLOT, means that were created in the UNIVARIATE procedure is imported as the dataset to be used in producing the interaction plot.

```
PROC SORT DATA=Data;
        BY IV time;
RUN;

PROC UNIVARIATE DATA=Data;
        BY IV time;
        VAR DV;
        OUTPUT OUT=means MEAN=DVmean;
RUN;

PROC SGPLOT DATA=means;
        SERIES y=DVmean x=time / GROUP=IV;
        TITLE 'Interaction plot';
RUN;
```

Where GROUP=IV is provided to look at each level of the independent variable and show the changes of the responses for each level of the independent variable over time using one noodle. Figure 3 provides an example of a spaghetti plot which illustrates how the mean of the response at each level of the predictor changes over time and how each level of the predictor changes compared to the other levels of the same predictor. This plot can be helpful in seeing the differences in the growth of each level of the factor of interest over time.

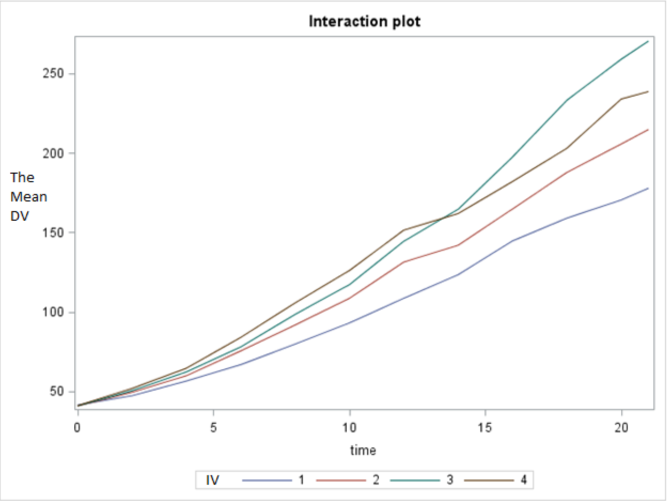

Figure 3. Interaction Plot

Plotting variograms, producing covariance, correlation and scatterplot matrices, and fitting simple models to the data are all helpful ways of studying and describing the data and their characteristics before moving toward more complicated longitudinal models.

CONCLUSION

Different options for modeling longitudinal responses were discussed above. Procedures developed within SAS such as PROC MIXED, PROC GENMOD, PROC GLIMMIX, PROC PHREQ, PROC MODEL, and PROC GEE can appropriately model correlated outcomes. PROC MIXED was also discussed for multilevel modeling where there exists correlation among observations for clustered data. Finally some simple graphing options were introduced as they can be extremely helpful in visualizing the trend of changes of different variables of interest over time which can be of interest. They can reveal helpful information especially regarding the type of correlation that exists among data which helps researchers in choosing the most appropriate inferential models to fit to the data in future.

The important point which was highlighted in this paper and needs to be considered by applied researchers and practitioners is the necessity of accounting for the correlation that exists among observations when modeling longitudinal or correlated data in general. Using appropriate correlated models which were discussed above will result in more informative and powerful models that are capable of answering research questions regarding the changes over time per subject or among different populations and clusters while correctly using the information regarding the autocorrelation of subjects in the process of fitting longitudinal and correlated models.

Author is currently working on extending this study in different directions such as longitudinal growth-curve modeling using HLM and developing easier ways of fitting GMM models to different types of response variables and covariates in the presence of complexities involved in the data such as missing observations.

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.

Bena, J., McIntyre, Sh. 2008. Survival Methods for Correlated Time-to-Event Data.

Diggle, P., Liang, K. Y., & Zeger, S. L. (1994). Longitudinal data analysis. New York: Oxford University Press, 5, 13.

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.

Hall, A. R. (2005). Generalized method of moments. Oxford: Oxford University Press.

Hansen, L. P. (1982), “Large-Sample Properties of Generalized Method of Moment Estimators,” Econometrica, 50, 1029–1054.

Hansen, L. P. (2007). Generalized Method of Moments Estimation. The New Palgrave Dictionary of Economics, ed. by Stephen Durlauf, and Lawrence Blume. Macmillan, forthcoming.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association, 72(358), 320-338.

Lai, T. L., & Small, D. (2007). Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(1), 79-99.

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

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

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.

SAS Institute Inc. 2014. SAS/ETS®13.2 User’s Guide. Cary, NC: SAS Institute Inc.

Zorn, C. J. (2001). Generalized estimating equation models for correlated data: A review with applications. American Journal of Political Science, 470-490.

RECOMMENDED READING

SAS Institute Inc. 2014. SAS/ETS®13.2 User’s Guide. Cary, NC: SAS Institute Inc.

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.

Ramezani, N. (2016). How to analyze correlated and longitudinal data?. In proceedings of the Western Users of SAS Software Conference. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

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

Niloofar Ramezani
University of Northern Colorado
Niloofar.ramezani@unco.edu