

Analysis of Repeated Measures and Longitudinal Data

Jane Pendergast, Dept. of Statistics, Univ. of Florida, Gainesville, FL

ABSTRACT

Repeated measures and longitudinal data require special attention because they involve correlated data. This correlation is induced when the primary sampling units are measured repeatedly over time or under different conditions. Examples of techniques applicable to normal and non-normal data using the SAS system will be presented. The focus of the presentation is an overview of methods which are available and help in choosing among them, rather than explaining how to properly use them and interpret the output. We take the perspective that the primary objectives of the analysis are to investigate trends over time and to determine if there is a relationship between the response variable and treatment groups or other covariates.

INTRODUCTION

The term "longitudinal data" implies that multiple measurements have been taken over time on a sampling unit. That sampling unit, or "subject", could be a person, plant, animal, piece of raw material, a machine, etc. In the discussion to follow, we will assume that we have a random sample of such subjects, each measured over time. We will not consider the situation where there is only one subject – that is, a single series of measurements taken over time, which could be analyzed using time series methods. Recent literature in this area often describes our setting as one in which we have independent "clusters" of observations. Within each cluster, there are multiple measurements which are assumed to be correlated. For example, in a setting where a random sample of n subjects are measured at five time points, we would have n independent clusters, and each cluster would be comprised of five correlated observations. Many techniques applied to longitudinal data can be applied more generally to other types of clustering. For example, multiple measurements taken at the same time on different parts of the subject, multiple subjects within the same family, or measurements physically close to each other (clustered in space) could also be considered correlated clusters of observations. However, this is not always the case. When the statistical method relies on being able to condition on the observed responses of other cluster members (e.g., the response conditional models and many of the methods applicable to spatial data), it would require conditioning on future observations to model characteristics of earlier observations, which would not make much sense.

If there are covariates, they could either vary from subject to subject, but not over time within one subject, or vary at different time points as well. For example, treatment group membership or gender may stay constant across time and would be considered a "between-cluster" or "between-subject" covariate, whereas the subject's weight could change from time point to time point and would be considered a "within-cluster" or "within-subject" covariate. When considering only longitudinal data, the terms "time-stationary" and "time-varying" are often used for "between-cluster" and "within-cluster", respectively. The goal of the analysis is to measure the relationship of these covariates upon the responses observed and to characterize trends over time.

Statistical issues which affect the covariance structure in the analysis of such data include:

1. recognizing the correlation structure within the data set and determining how best to model this correlation;
2. determining whether the observations within the cluster were taken in a fixed order (e.g., at given time points) or if they can be in random order – for example, multiple measurements taken on the same subject under different conditions at (approximately) the same time;
3. determining whether the number of observations within a cluster can vary; and
4. determining whether the multiple observations over time were taken at the same set of time points (a pre-specified schedule of time points) or allowed to vary from subject to subject.

Failing to account for the correlation in a meaningful way will result in standard errors of the model parameters which are incorrect, problems with the coverage probabilities of interval estimates, inaccuracy of the stated levels of hypothesis tests, etc. Advantages of longitudinal studies include the ability to distinguish between variation in one subject across time from variation among different subjects and probably greater efficiency in the design and in the estimation. It may be cheaper to take multiple measurements on the same subject under different conditions than to recruit more subjects. By controlling for subject to subject variation, we can get a more sensitive estimate of how responses are likely to change over time.

EXAMPLE DATA SET: HEALTHY KIDS MEDICAL INSURANCE PROGRAM

The State of Florida has implemented, in some counties, a program called the Healthy Kids Program through which children can obtain health insurance coverage if they are not Medicaid-eligible. The county school district is used as a grouping mechanism to spread the financial risk across a large number of children, rather than the traditional mode whereby a risk pool is created via the parent's employer. Costs are kept relatively low by restricting the coverage only to children, who tend to be reasonably healthy, and not to their parents. The cost of providing care is shared among the family, the State of Florida, organizations within the County, and the managed care organization (MCO) providing the care. Families are required to pay a monthly premium, which can be subsidized at various levels depending on the financial resources of the family. Managed care organizations in the county compete for a contract awarded by the State with a prespecified benefit package to provide the care. They share in the financial support by discounting their usual capitation rate, which is a fixed fee per child per month assessed to provide the health care.

The main goal of the Healthy Kids Program is to remove the financial barrier to obtaining needed health care, especially preventive care, which is expected to result in a healthier population of children and overall lower health care costs over time. A primary question of interest is whether or not lessening (or removing) the financial barrier affects the likelihood or level of use. How do the use patterns change over time? Are other covariates useful in predicting such use? We will consider the outcome variable to be the rate of use, as measured by the number of days on which the child used the services divided by the number of months the child was enrolled in the program. Possible covariates

include the subsidy level (called plan level), the age of the child at the beginning of his/her enrollment in the program, and the number of months the child was enrolled in the program each year.

TRADITIONAL METHOD OF ANALYSIS FOR REPEATED MEASURES DATA

One approach to analyzing such use rates over time would be to use repeated measures ANOVA analysis. We would like to model the yearly use rates for a particular child as a function of the child's gender, age at the beginning of the time frame, the year the rate was calculated, the Plan level, and the number of months the child was enrolled. However, there is a problem. Repeated measures ANOVA does not allow covariates to change across time, and since the number of months the child was enrolled and the Plan level could change every year. Sometimes it makes sense to use some function of these time-varying covariates, such as the total number of months enrolled over all years (or average per year) and the plan level at the beginning of the time frame or the modal plan level across time. For the purpose of this example, we will use the average number of months enrolled per year and the initial plan level. Another potential problem with this analysis is that it requires all children to have measurements at every time point. In many longitudinal data sets, not all subjects can be measured at each and every time point. Any information contained in those observations with some missing values is lost, since they must be eliminated from the analysis. In our data set, there are records on 3406 unique children, but only 2308 were enrolled for the entire five years.

This methodology is based on the assumption of normally distributed responses, and the use rate measures have a distribution that tends to be fairly skewed to the right. The log transformation (adding a small fudge factor since zero visits in a year is possible) helps to stabilize the variance, makes the response distribution more symmetric and shall be used here.

SAS Statements for Repeated Measures ANOVA Using GLM

```
PROC GLM DATA=REPMEAS;
  CLASS SEX PL9192;
  MODEL LNRATE91-LNRATE95 = SEX PL9192 AVE_MNS
  AGE AGE*AGE AGE*AGE*AGE;
  CONTRAST 'PLAN 12 -PLAN 34' PL9192 1 1 -1 -1;
  TEST 'OVERALL AGE' AGE = 0, AGE*AGE=0,
  AGE*AGE*AGE=0;
  LSMEANS PL9192 SEX / TDIFF;
  REPEATED YEAR 5 (1 2 3 4 5) PROFILE / SUMMARY
  PRINTM PRINTE;
  TITLE1 "REPEATED MEASURES ANOVA";
```

The child's gender and plan level at the beginning of the period are entered into the model as CLASS variables, but age of the child (AGE) and average months enrolled per year (AVE_MNS) are treated as continuous covariates. Because the expected health care use of a child varies considerably depending on his/her age, a cubic polynomial in age was used in the model.

In order for the univariate *F* tests which result from this analysis to be valid, there is an assumption on the variance/covariance matrix of the repeated measurements over time. The vector of responses over time for any subject (i.e., SQRATE91 - SQRATE95), must have and "TYPE H" matrix, meaning that the variance/covariance matrix of a set of orthogonal contrasts in these response variables must be compound symmetric - that is, equal variances and equal covariances. If SQRATE91-SQRATE95 have homogeneous variances and the same pairwise correlation between any two time

points, then this condition will be met. However, it can be met in other ways as well. You can check the validity of this sphericity assumption by using the PRINTE option in the REPEATED statement. The Greenhouse-Geisser (1959) and Huynh-Feldt (1970) adjustments to the degrees of freedom of the *F* tests can be applied to obtain adjusted *p* values, if sphericity is rejected. However, these adjustments are limited in their usefulness, and a strong rejection of sphericity would be good reason to seek a different approach to the analysis.

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA) APPROACH

Use of a multivariate analysis of variance (MANOVA) approach to the problem eliminates the need for a sphericity assumption. Here, the multiple responses over time are treated as simply a random vector, with an arbitrary variance/covariance matrix. This expectation of this vector (i.e., the vector of mean responses at each time point) is modeled in terms of the between-cluster (or time-stationary) factors and covariates. The use of the REPEATED statement in PROC GLM as illustrated above will generate the multivariate tests of AGE, AGE², AGE³, PL9192 and AVE_MNS, testing whether the mean response vector over time is different at different ages, at different plan levels, etc. Comparisons among time points are accomplished using transformations of the response vector. Some common transformations such as a set of orthogonal polynomials, Helmert contrasts, differences between adjacent time points, differences between one point and the mean of the rest and differences from a control group (such as baseline) are easily accomplished using keywords in the REPEATED statement (i.e., POLYNOMIAL, HELMERT, PROFILE, MEAN and CONTRAST, respectively), but user defined transformations are possible as well.

Like the univariate repeated measures approach, the hypothesis tests following the MANOVA model are based upon the assumption of normality. Here we assume that the response vectors (SQRATE91-SQRATE95) have a multivariate normal distribution with a common, but arbitrary, variance/covariance matrix. Since you would expect measurements taken close together in time to be more highly correlated than those far apart, the assumption of a common covariance matrix would generally imply that all subjects were measured at the same set of time points. As the number of time points increases, the number of parameters to be estimated increases rapidly, thus requiring greater numbers of subjects to provide good estimates and reasonable power. Missing data, especially at the end of the time frame, is quite common in longitudinal data, as is time points varying from subject to subject. If choosing between univariate repeated measures ANOVA and the corresponding MANOVA models, one would have to balance the gain to be made with the added flexibility of the covariance structure with the cost of additional subjects to maintain sufficient power. The other assumptions and limitations are the same in both approaches.

MIXED MODEL ANALYSIS

The mixed model approach for longitudinal data attacks the correlation issue more directly by including random parameters in the model which capture the correlation structure both between and within subjects. In order to understand this type of model, it helps to start by reviewing the usual linear model for normally distributed data (i.e., a regression or analysis of variance model). In matrix notation, the usual linear model has the form:

$$y = X\beta + \epsilon$$

where

$\mathbf{Y}$ is the $n \times 1$ vector of responses on n independent subjects
 $\mathbf{X}$ is the $n \times p$ matrix of values of the factors and covariates
 β is the $p \times 1$ vector of regression parameters, and
 ϵ is the $n \times 1$ vector of independent errors.

We assume the errors have a common mean (zero) and common variance (σ^2) and usually assume that they are normally distributed as well (i.e., $\epsilon \sim N(0, \sigma^2 I)$). With or without the normality assumption, the ordinary least squares (OLS) estimate of the β parameter is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

which is obtained by solving the normal equations:

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}.$$

If we add the normality assumption, this estimator is also the maximum likelihood estimator, but even without that assumption, it has good statistical properties as the OLS estimator. Note that the β parameters are all considered fixed effects in this model.

The mixed model builds on this foundation and adds random effects to the model. It has the form:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon$$

where $\mathbf{u}$ is a vector of random parameters, which is assumed to have a multivariate normal distribution with mean 0 and variance/covariance matrix $\mathbf{G}$ (i.e., $MVN(0, \mathbf{G})$), and $\mathbf{Z}$ is a design matrix for those parameters. One further distinction is that we now generalize the possible covariance structures for the errors and assume ϵ is $MVN(0, \mathbf{R})$, where $\mathbf{R}$ is any valid covariance matrix, not just $\sigma^2 I$. This structure implies that the covariance matrix of $\mathbf{Y}$ is $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$, a mix of correlation coming from the random effects and from the individual error terms ($\mathbf{Y} \sim MVN(\mathbf{X}\beta, \mathbf{V})$). While the unconditional mean of $\mathbf{Y}$ is $\mathbf{X}\beta$, the conditional mean given the random effects is $E(\mathbf{Y} | \mathbf{u}) = \mathbf{X}\beta + \mathbf{Z}\mathbf{u}$. The analyst can specify the form of $\mathbf{G}$ in the RANDOM statement, and the form of $\mathbf{R}$ in the REPEATED statement. Some of the different types of covariance structures which can be specified are presented in Table 1. Since these matrices are often a function of unknown parameters, they must be estimated. SAS uses maximum likelihood estimation (ML) or restricted maximum likelihood (REML) estimation to obtain such estimates of $\mathbf{G}$, $\mathbf{R}$, and $\mathbf{V}$. Once obtained, they are plugged into the mixed model equations (generalized least squares):

$$\begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Z} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z} + \hat{\mathbf{G}}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Y} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Y} \end{bmatrix}$$

with resulting solution:

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1} \mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y} \\ \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta}) \end{bmatrix}.$$

The resulting covariance matrix of $\hat{\beta}$ has the form

$$\mathbf{V}(\hat{\beta}) = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}.$$

Table 1: Some Forms of Covariance Matrices in PROC MIXED

Description	Form
Independent error	$\Sigma = \sigma^2 I$
Compound Symmetry (also known as Exchangeable)	$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$ $\sigma_{ij} = \begin{cases} \sigma^2 & i = j \\ \sigma^2 \rho & i \neq j \end{cases}$
Autoregressive of order 1	$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$ $\sigma_{ij} = \sigma^2 \rho^{ i-j }$
Autoregressive of order 1 with heterogeneous variances	$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1 \sigma_2 \rho & \sigma_1 \sigma_3 \rho^2 \\ \sigma_1 \sigma_2 \rho & \sigma_2^2 & \sigma_2 \sigma_3 \rho \\ \sigma_1 \sigma_3 \rho^2 & \sigma_2 \sigma_3 \rho & \sigma_3^2 \end{bmatrix}$ $\sigma_{ij} = \sigma_i \sigma_j \rho^{ i-j }$
Unstructured	$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$

Example of Fitting a Linear Mixed Model Using the Healthy Kids Data

While the mixed model approach also assumes normality, it is more general than the repeated measures ANOVA or the multivariate MANOVA in that it can accommodate some missing data, deals with random effects in the model in an appropriate manner, and allows a variety of types of covariance structures. An example using PROC MIXED to model the use rate is presented below. To capture the correlation over time, we assume an AR(1) error structure for the $\mathbf{R}$ matrix. To illustrate how other random effects ($\mathbf{u}$) can be included in the model as well, we include a random parameter for each subject (i.e., a random intercept in the model) which will make the $\mathbf{G}$ matrix a scalar (σ^2). Options to print out the estimated $\mathbf{R}$, $\mathbf{G}$ and resulting $\mathbf{V}$ covariance matrices are included in the RANDOM and REPEATED statements. Note that although the REPEATED statement can be found in both GLM and MIXED and plays a similar role, the syntax is very different in the two procedures.

```
PROC MIXED DATA = LONG METHOD = ML;
  CLASS YEAR PL9192 SEX /*SUBJ*/;
  MODEL RATE = SEX PL9192 YEAR AGE
    AGE*AGE AGE*AGE*AGE /
    SOLUTION CHISQ;
  LSMEANS YEAR PL9192 / PDIF;
  CONTRAST 'OVERALL AGE' AGE 1,
    AGE*AGE 1, AGE*AGE*AGE 1;
  ESTIMATE '1991, PL1' INTERCEPT 1
    AGE 0 AGE*AGE 0 AGE*AGE*AGE 0
    YEAR 1 0 0 0
    PL9192 1 0 0 0 / CL;
  RANDOM INTERCEPT / SUBJECT = SUBJ G
    V V CORR /* SOLUTION */;
  REPEATED / TYPE = AR(1) R SUBJECT = SUBJ;
```

```
TITLE1 'Linear Mixed Model with Random
      Intercept and AR(1) Error';
title2 'GLMMIX Macro Output';
title3 'Includes Age, Age2, Age3 Centered At
10';
title4 'Plan Level Held Stationary In Time';
```

The SUBJECT= option in the RANDOM statement instructs SAS that this variable identifies the different independent units (clusters). In other words, it tells SAS when a different value(s) of the random effect(s) (intercept here) should be used in the model. The use of the same option (SUBJECT =) in the REPEATED statement is used to speed up calculations. The SOLUTION option in the RANDOM statement will generate output of the estimates of the random effects along with associated standard errors. The LSMEANS, CONTRAST and ESTIMATE statements are included to illustrate how they can be used in MIXED to provide further detailed analyses. The publication, *SAS System for Mixed Models* by Littell et al. (1996) provides comprehensive coverage of how to use PROC MIXED for various types of analyses.

GENERALIZED LINEAR MODELS (GLMs)

The familiar linear model $Y = X\beta + \varepsilon$ was described above assuming normally distributed errors. In practice, we often encounter data where the normality assumption is not reasonable – dichotomous responses, count data, skewed data, etc. Sometimes the response is known to be constrained within limits or the variance is expected to change with the mean. Classical approaches to handling such problems are to use transformations on the response (e.g., the variance stabilizing arcsine transformation for binary data), weighted or generalized least squares, transformations on the predictors, or some combination of the above.

Nelder and Wedderburn (1972) presented an alternative to the classical linear model for independent observations (without random effects) known as the *generalized linear model (GLM)*. They extended the normal theory linear model by allowing more general functional relationships between the mean response and the linear combination $X\beta$, and by accommodating a more general class of error distributions. The GLM is described by three components:

1. a *random* component - which describes the distribution of the response variable which can be continuous or categorical;
2. a *systematic* component - which describes how covariates enter into the model for the mean; and
3. a *link* between the random and systematic component, which describes the functional relationship between the mean response and the systematic component.

and can be fit in SAS using PROC GENMOD.

Random Component

GLM models can be used to model a random sample of univariate responses ($Y_1, Y_2, \dots, Y_n$) from a distribution within the exponential family. This is a broad class which includes familiar distributions such as the normal (Gaussian), Binomial, Poisson, Gamma and Inverse Gaussian, all of which are easily specified in SAS. The *canonical parameter* is a term for a natural parameter within the distribution and is some function of $E(Y)$.

Systematic Component

The term $X\beta$ which contains the factor/covariate values and the fixed regression parameters is called the *linear predictor* in the terminology of GLMs. It is often denoted by the vector η .

Link Function

The relationship between the i th component of the linear predictor and mean of the i th response ($E(Y_i) = \mu_i$) is described by the link function, $g(\cdot)$. That is, $\eta_i = X_i\beta = g(\mu_i)$. Any monotonic differentiable function is an acceptable link, but the particular one which relates the mean to the canonical parameter is called the *canonical link* and is often used in practice. The canonical link is the default in PROC GENMOD, unless the analyst specifies otherwise.

Some distributions in the exponential family have a scale parameter, which is generally considered a nuisance parameter in this setting. For example, the normal (Gaussian) distribution is a member of the exponential family, with the variance parameter considered a nuisance scale parameter. In some distributions, the scale parameter is known, or known up to some constant. Often it will have the form ϕ/w , where w is a known weight.

Maximum likelihood estimation is feasible in this setting and is accomplished with an iterative computational scheme. If we back off the full distributional assumption and assume only a functional relationship between the mean and variance, we can base our estimation on a different type of likelihood, called *quasi-likelihood*. Quasi-likelihood estimators have similar asymptotic properties, but differ in that they only require assumptions on the first two moments of the underlying distribution.

Example of Fitting a GLM Using the Healthy Kids Data

In the Healthy Kids data, we are trying to model the children's use of the health care services over several years. To illustrate the use of a GLM model, we will only use the data from the first year of the program, 1991. That way, there will only be one observation per year per child and thus the assumption of independent responses will be reasonable. If we simply count the number of uses per year (TOTUSE91) rather than create a use rate per month enrolled, we could reasonably assume that these counts had a Poisson distribution. The canonical link for the Poisson distribution is the log link, which will be used here. As before, we will use the age and sex of the child, the plan level, and the number of months enrolled in the program as covariates to model the mean response, so they will be part of the linear predictor.

The SAS program statements needed to fit this model, again assuming a cubic polynomial in AGE, are given below.

```
PROC GENMOD DATA=REPMEAS;
  CLASS SEX PL9192;
  MODEL TOTUSE91 = SEX PL9192 MNS9192
    AGE AGE*AGE AGE*AGE*AGE / DIST = POI
    LINK= LOG;
  CONTRAST 'PLAN 1 - PLAN 2' PL9192 1 -1 0 0;
  CONTRAST 'PLAN 1 - PLAN 3' PL9192 1 0 -1 0;
  CONTRAST 'PLAN 1 - PLAN 4' PL9192 1 0 0 -1;
  CONTRAST 'PLAN 2 - PLAN 3' PL9192 0 1 -1 0;
  CONTRAST 'PLAN 2 - PLAN 4' PL9192 0 1 0 -1;
  CONTRAST 'PLAN 3 - PLAN 4' PL9192 0 0 1 -1;
  TITLE1 "GLM Model - Health Kids 1991";
```

```
TITLE2 "Poisson Dist, Log Link (Canonical
Link)";
TITLE3 "Response Variable Is The Number Of
Health Care Visits";
```

The DIST=POI and LINK=LOG options of the MODEL statement specify the Poisson distribution (or more generally, the same mean-variance relationship as is seen in the Poisson distribution) and the log link. If we preferred to model the rate of use per month of enrollment rather than the count of the number of "visits", the log of the denominator of the rate (i.e., LOGMNS = log(MNS9192)) would be the OFFSET variable. The OFFSET option in the MODEL statement is used to identify this variable, which cannot be included as a covariate as well. The revised model statement has the form:

```
MODEL TOTUSE91 = SEX PL9192 AGE AGE*AGE
AGE*AGE*AGE / DIST = POI LINK= LOG
OFFSET = LOGMNS;

TITLE4 "Log(MNS9192) Used As The Offset Variable
To Model The Monthly Use Rate";
```

Maximum likelihood estimation is used to obtain parameter estimates, and a measure called *deviance* is the basis for assessments of how well the model fit the data and to compare different models. Due to space limitations, we will not discuss how to assess the adequacy of the model fit using statistics based on the deviance measure.

GENERALIZED ESTIMATING EQUATIONS (GEEs)

The above example using a GLM model only considered the use rates for one year, which sidestepped the real problem in longitudinal data – that of correlated observations over time. However, it did address the issue of response distributions other than the normal distribution. What we still need is a way to incorporate the correlation over time into an analysis which can accommodate non-normally distributed responses. To do a likelihood analysis, we would have to characterize the multivariate distributions of the multiple responses. However, when we move outside the familiar Gaussian (normal) distribution, we find that not all univariate distributions have multivariate analogues that are accessible for analysis purposes. We also find that when we move outside the normal-theory framework, it becomes important to distinguish between two different types of models – *subject-specific* and *marginal* (or *population-averaged*) models, because they address different types of questions. Whereas there is no difference in a standard regression analysis between the average of the predicted values and the predicted value at the mean values of all the covariates, the same would not be true in logistic regression. In that setting, the logit computed on the mean response over subjects would not be the same as the average of the subject-specific logits. This underscores the role of the link function in GLMs, when the link is not the identity function (as it is in regression analysis) and helps us distinguish between subject-specific and marginal models.

Subject-Specific versus Marginal Models

The linear mixed model presented above provided a means by which changes in the response over time can be modeled in terms of covariates/factors, which account for random measurement error and within-subject correlation and variability. The random effects in the model drive that within-subject correlation. Since these random effects parameters can vary from subject to subject, (or more generally, from cluster to cluster), such models are known as

"subject-specific" or "cluster-specific" models. That is, they model what is happening within one subject or cluster by using subject-specific (cluster-specific) parameters. These models are appropriate when we want to interpret how differing covariate values are likely to affect the response of a subject, conditional on all other covariates. We first model what happens within an individual subject, then under the assumption that the beta parameter for this covariate is the same for all subjects, we pool information across subjects to estimate it. In contrast, models which focus on how the covariates are related to marginal mean responses after averaging (in some sense) over individual subjects are known as *marginal* or *population-averaged* models. The difference in focus of these two types of models reflects the differences in the type of questions each can address and the corresponding interpretations of model parameters. Further discussions of the difference between these two can be found in Zeger et al. (1988), Neuhaus et al. (1991), and Pendergast et al. (1996), among others.

Generalized Estimating Equations (GEEs)

When the analyst is primarily interested in marginal models, the correlation or association parameters expressing relationships over time are really not of scientific interest, but rather are nuisance parameters. We must account for them in order to make valid inferences on the marginal mean parameters (i.e., those describing the means at each time point), but they are not the main focus of the investigation. Liang and Zeger (1986) and Zeger and Liang (1986) presented an estimation method for incorporating within-cluster correlation in marginal models, treating these correlation parameters as nuisance parameters. They started by assuming GLMs for the marginal means, developing the estimating equations which would be appropriate if all the observations across time were independent, and then generalized those estimating equations to incorporate an assumed or "working" correlation structure. While this approach does not produce maximum likelihood estimates, the GEE estimates have known asymptotic distributions and have performed well in practice. Failure to account for the correlation across time would result in estimators which were inefficient and had poor estimates of their standard errors. They later presented this approach simply assuming a mean-variance relationship, which results in quasi-likelihood estimating equations.

Much of the literature on GEE estimation discusses two different types of standard errors of the beta parameters. The first, called the *naïve* estimator, constrains the covariance matrix of the $\hat{\beta}$ estimates to have the functional form of the marginal variances implied by the specified distribution (e.g., $p(1-p)/n$ for a binomial proportion) and the working correlation matrix. The analyst may prefer to let the data have more influence over the structure of this covariance matrix, and that can be accomplished by using a form of the variance estimate called the *robust* estimator. Many researchers in this area prefer to use the robust estimator, if the number of clusters is reasonably large relative to the cluster size.

Besides accommodating non-normal data, this estimation procedure can handle both within- and between-cluster covariates and a certain type of missing data. The responses can be missing either at individual time points or as a truncation of the series, as long as these data can be considered missing completely at random (MCAR), in the sense of Little and Rubin (1987). A good introductory reference to this methodology is a recent book by Diggle, Liang and Zeger (1994).

Example of GEE Estimation Using the Healthy Kids Data

The GEE approach to estimation is available in the 6.12 release of PROC GENMOD in SAS. The regression model of the marginal

mean is specified the same as for ordinary GLMs in the MODEL statement, and the working correlation structure is specified in the REPEATED statement.

To illustrate the use of the GEE estimation in GENMOD, we will again use the Healthy Kids data. However, we need to restructure the data set so that there is one observation per time point, rather than one observation per child. A new variable, called YEAR, is created to identify which year, TOTUSE holds the number of health care visits that year, and MNS holds the number of months the child was enrolled in that program during that year. For simplicity, we will maintain PL9192 as a between-child covariate, rather than allowing the plan level to change from year to year. The age of the child at the beginning of the time frame and the gender of the child are retained as between-child covariates. The syntax of the REPEATED statement is different than in PROC GLM or PROC MIXED¹ and controls some of the options, so we will describe it in more detail here. It has the form:

```
REPEATED SUBJECT = subject-effect
</options>;
```

where "subject-effect" is the name of the variable which uniquely identifies the subject who has multiple measures (i.e., the cluster identifier). It can be specified as a single variable, an interaction effect, a nested effect or combinations. Any variable used to identify the subject uniquely must be included in the CLASS statement.

The options control how the model is fit and what output is produced. A key option is the one used to specify the working correlation structure, and can be specified either using the syntax TYPE = keyword or CORR = keyword. Table 2 presents the various working correlation structures available in GENMOD.

Table 2. Working Correlation Structures Available for GEE Estimation in GENMOD

Keyword	Working Correlation Matrix
AR AR(1)	Autoregressive of order one
EXCH	Exchangeable (Compound Symmetry)
IND	Independence
MDEP(number)	M-dependent, with m = number
MODGEEML	Modified GEE with unstructured
ML	correlation and maximum likelihood estimation for normal data
UNSTR UN	Unstructured
USER(matrix)	Fixed, user-specified correlation matrix.

The 1-dependence structure implies that all elements of the first off-diagonal in the correlation matrix are the same common correlation parameter, but all other correlation parameters are zero. This would be a reasonable model if one thought that adjacent time points would be correlated and a common value of that correlation was likely, but non-adjacent time points were not correlated. The 2-dependence structure builds on 1-dependence by also including a common (but potentially different) correlation parameter on the second off-diagonal, and padding the rest of the matrix with zeroes. Here, time points one and two steps apart would be assumed correlated, but not those further apart. In the general m -dependence case, time points m steps apart and closer are

assumed correlated, with a common correlation coefficient for each distance apart, and those further apart are assumed uncorrelated.

Options of the REPEATED statement which control output allow you to print out the estimated working correlation matrix (CORRW), both the naive and robust estimated covariance matrix of the beta coefficients (COVB), and/or a printed table of response values, predicted values, confidence limits on the predicted values, the linear predictor and residuals for each observation (OBSTATS). If you prefer to use the naive (model based) estimator of variance in the analysis of parameter values table rather than the robust (which is the default), this can be accomplished using the MODELSE option. Initial starting values (INITIAL = number), the maximum number of iterations (MAXIT=number), convergence criteria (CONVERGE=number), and/or a fixed value of the intercept (INTERCEPT = number) can be specified as options in the REPEATED statement (in conjunction with the NOINT option of the MODEL statement).

The following is an example of using GENMOD to fit a marginal model to the rate of use over time, assuming the number of uses per year had a Poisson distribution. In addition, we assume that the within-subject correlation over time has the AR(1) structure. Contrasts between marginal means at different plan levels is accomplished via the use of the CONTRAST statements. Age of the child enters into the model as a cubic polynomial, to characterize the changing patterns of use as a child matures.

```
PROC GENMOD DATA=LONG;
  CLASS SEX PL9192 SUBJ;
  MODEL TOTUSE = SEX PL9192 AGE AGE*AGE
    AGE*AGE*AGE / DIST = POI
  LINK= LOG OFFSET = LOGMNS;
  CONTRAST 'PLAN 1 - PLAN 2' PL9192 1 -1 0 0 ;
  CONTRAST 'PLAN 1 - PLAN 3' PL9192 1 0 -1 0 ;
  CONTRAST 'PLAN 1 - PLAN 4' PL9192 1 0 0 -1 ;
  CONTRAST 'PLAN 2 - PLAN 3' PL9192 0 1 -1 0 ;
  CONTRAST 'PLAN 2 - PLAN 4' PL9192 0 1 0 -1 ;
  CONTRAST 'PLAN 3 - PLAN 4' PL9192 0 0 1 -1 ;
  REPEATED SUBJECT=SUBJ / TYPE=AR(1) CORRW COVB;
```

GENERALIZED LINEAR MIXED MODELS (GLMMs)

Two key features of the linear mixed model described above are the fact that the mean response is equal to the linear predictor -- that is, the identity link -- and that the random components are assumed to have multivariate normal distributions. In the spirit of generalized linear models, the *generalized linear mixed model (GLMM)* allows link functions other than the identity as well as the broader class of distributions in the exponential family (random component). While it is possible to obtain maximum likelihood estimates within the GLM framework, the theory underlying the GLMMs backs off true maximum likelihood and relies on other types of likelihood functions instead.

In the generalized linear mixed model, the conditional mean $E(Y|u)$ plays the same role as the unconditional mean in the GLM. This conditional mean is related to the linear predictor via the link function $g(\cdot)$:

$$\eta = g(\mu|u) = X\beta + Zu$$

with conditional variance

$$\text{var}(Y|u) = \text{var}(e) = R.$$

¹ Documentation on the syntax to implement GEE estimation and an example can be obtained from the SAS Institute on the WWW at the URL <http://www.sas.com/techsup/download/stat/gee.txt>.

The analyst chooses an appropriate link function and error distribution for the conditional distribution.

SAS GLIMMIX Macro

Estimation methods and theory underlying this type of model are still being developed and assessed, and software to fit such models is not readily available in common statistical packages. However, the SAS Institute has a macro, called *glimmix*, written by Russ Wolfinger and available in an "as is" testing mode, which can fit such models.² We illustrate its use in the Healthy Kids data example below, where the data set has been structured as one observation per time point, rather than one observation per child.

```
%INC 'C:\SAS\MACROS\GLMM612.SAS' / NOSOURCE;

%GLIMMIX(DATA=SCDATA.LONG,
  PROCOPT=METHOD=REML,
  STMTS = %STR(
    CLASS YEAR PL9192 SEX SUBJ;
    MODEL TOTUSE = SEX PL9192 YEAR AGE
      AGE*AGE AGE*AGE*AGE;
    /*SOLUTION CHISQ;*/
    LSMEANS YEAR PL9192 / PDIFF;
    ESTIMATE '1991, PL1' INTERCEPT 1
      AGE 0 AGE*AGE 0 AGE*AGE*AGE 0
      YEAR 1 0 0 0 0
      PL9192 1 0 0 0 / CL;
    CONTRAST 'OVERALL AGE' AGE 1,
      AGE*AGE 1, AGE*AGE*AGE 1;
    REPEATED / TYPE = AR(1) SUBJECT = SUBJ R;),
  ERROR = POISSON,
  OFFSET = LOGMNS,
  LINK = LOG,
  MAXIT = 50,
  TITLE = 'Generalized Linear Mixed Model with
AR(1) Error');
title2 'GLMM612.SAS (GLIMMIX) Macro';
title3 'Includes AGE, AGE2, AGE3 Centered At
10';
title4 'Plan Level Held Stationary In Time';
```

The basic idea is that this macro iteratively calls PROC MIXED until convergence. The PROCOPT= statement allows you to list options available in PROC MIXED, such as whether maximum likelihood (ML) or REML estimation is preferred. While the ML option may be specified, it does not actually produce maximum likelihood estimates. The macro uses iteratively reweighted likelihoods to fit the model as discussed in Wolfinger and O'Connell (1993). The macro can use a *pseudo-likelihood (PL)* or a *restricted/residual pseudo-likelihood (REPL)* (the default) to estimate the parameters of the model. The PL approach assumes that the scale parameter in the GLM is known, whereas the REPL approach assumes it unknown. The PL method parallels a maximum likelihood approach and the REPL method obtains

REML-type estimates. Breslow and Clayton (1993) take a similar approach, but use a different type of likelihood.

The STMTS = parameter allows the user to string out PROC MIXED statements, using the syntax specified for that procedure. This is the place where the covariance structures of R and G are defined, as well as the linear predictor of the conditional generalized linear model. Any of the following statements may be included: CLASS, MODEL, RANDOM, REPEATED, PARMS, ID, CONTRAST, ESTIMATE, and LSMEANS. The options which specify components of the generalized linear model, such as the link function and the error distribution, come afterwards. You may specify preprogrammed links, such as the logit, log, identity, reciprocal, and power(-2) links, which are the defaults (canonical links) for the Binomial, Poisson, Normal, Gamma, and Inverse Gaussian distributions, respectively. However, the user can define his/her own link function if needed. Parameters which control the printed output, initialize starting values, control the iterations and/or output information into a file are available as well. See the *glmm612.sas* and *glmm611.sas* files for further detail on the options available.

CONCLUSION

When faced with the task of analyzing longitudinal data, the investigator must make a series of decisions which will lead him/her to an appropriate method. Can the responses be assumed to be normally distributed, or would some other distribution be a more reasonable assumption? What questions need to be addressed? Should a marginal model or subject-specific model be used to address them? Are there a lot of missing data? What type of covariance structure over time would be reasonable? Do I need to accommodate time-varying covariates and/or random effects? Table 3 below summarizes some of the differences and similarities in the statistical approaches discussed in this paper. While none of these approaches would be preferred uniformly over the others, there are some general comments which could be made. If the correlation over time is expected to be low, then the specified structure of the covariance matrix will not play as large a role as it does with strong correlation. If there is strong correlation, gains in efficiency can be high when modeling this structure with fewer parameters (e.g., AR(1) process rather than a full arbitrary covariance matrix as in the MANOVA approach. When the data are non-normal, either a direct approach (e.g., assuming a binomial error distribution) or a transformation (e.g., arcsine) can be used. Parameter interpretation may be easier with the direct approach. Eliminating all clusters with any missing observations can change the sample size dramatically and potentially bias the results toward those cases which (for whatever reason) were able to be measured consistently. However, it should be noted that in many longitudinal data settings, the pattern of missingness is not MCAR, but rather resulting from a process related to the response which would have been observed. The ability to handle time-varying covariates is a big plus in many situations.

² The macro and supporting information for fitting a generalized linear mixed model can be obtained from the SAS Institute on the WWW at URL <http://www.sas.com/techsup/download/stat>. Download the files *glmm611.sas* and *glmm611e.sas* for Release 6.11 and earlier; download *glmm612.sas* and *glmm612e.sas* for use with Release 6.12.

Table 3. Comparison of Statistical Approaches to Longitudinal Data

	Repeated Measures ANOVA	Multivariate ANOVA	Linear Mixed Model	Generalized Estimating Equations (GEE)	Generalized Linear Mixed Model
SAS Procedure.	GLM	GLM	MIXED	GENMOD	GLIMMIX Macro
Accommodate some types of missing data	No	No	Yes	Yes	Yes
Allow time points to change from subject to subject	No	No	Yes - depending on covariance structure	Yes - depending on covariance structure	Yes - depending on covariance structure
Assume Normality	Yes	Yes	Yes	No	No
Assume Identity Link Function	Yes	Yes	Yes	No	No
Subject-Specific Model	Yes	Yes	Yes	No	Yes
Marginal Model	Yes	Yes	No	Yes	No
Allow Time-varying Covariates	No	No	Yes	Yes	Yes
Allow Random Effects in Linear Predictor	No	No	Yes	No	Yes
Flexibility in Specifying Covariance Matrix	No - assume Type H matrix	No - assume unstructured matrix	Yes	Yes - as a working covariance matrix only	Yes
Treat Correlation Parameters as nuisance	No	No	No	Yes	No
ML or REML estimation	Yes	Yes	Yes	No	No

REFERENCES

- Breslow, N.E. and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models", *Journal of the American Statistical Association*, 88, 9-25.
- Diggle, P. J., Liang K.-Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Clarendon Press, Oxford, 253 pp.
- Greenhouse, S. W. and Geisser, S. (1959), "On Methods in the Analysis of Profile Data", *Psychometrika*, 32, 95-112.
- Huynh, H. AND Feldt, L. S. (1970). "Conditions under Which Means Square Ratios in Repeated Measurements Designs have Exact F-Distributions", *Journal of the American Statistical Association*, 65, 1582-1589.
- Liang, K.-Y. and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models", *Biometrika*, 73(1), 13-22.
- Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996), *SAS System for Mixed Models*, Cary, NC: SAS Institute Inc. , 633pp.
- Littell, R. J. A. and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.
- Nelder, J. A. and Wedderburn, R. W. M. (1972), "Generalized Linear Models", *Journal of the Royal Statistical Society, Series A*, 135, 370-384.
- Neuhaus, J.M, Kalbfleishch, J. D., and Hauck, W. W. (1991), "A Comparison of cluster-specific and population-averaged Approaches for Analyzing Correlated Binary Data", *International Statistical Review*, 59(1), 25-35.
- Pendergast, J.F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., and Fisher, M. R. (1996). "A Survey of Methods for Analyzing Clustered Binary Response Data", *International Statistical Review*, 64, 89-119.
- Wolfinger, R. and O'Conner, M. (1993), "Generalized Linear Models: A Pseudo-likelihood Approach", *Journal of Statistical Computing and Simulation*, 48, 23-243.
- Zeger, S. L. and Liang, K.-Y. (1986), "Longitudinal data analysis for discrete and continuous outcomes", *Biometrics*, 42, 121-130.
- Zeger, S. L., Liang, K.-Y., and Albert, P. (1988), "Models for Longitudinal Data, A Generalized Estimating Equation Approach", *Biometrics*, 44, 1049-1060.

CONTACT INFORMATION

Jane Pendergast, Ph.D., Division of Biostatistics, Box 100212
JHMC, University of Florida, Gainesville, FL 32610-0212;
pender@stat.ufl.edu