1 Introduction

Correlated response data, either discrete (nominal, ordinal, counts), continuous or a combination thereof, occur in numerous disciplines and more often than not, require the use of statistical models that are nonlinear in the parameters of interest. In this chapter we briefly describe the different types of correlated response data one encounters in practice as well as the types of explanatory variables and models used to analyze such data.

1.1 Correlated response data

Within SAS, there are various models, methods and procedures that are available for analyzing any of four different types of correlated response data: (1) repeated measurements including longitudinal data, (2) clustered data, (3) spatially correlated data, and (4) multivariate data. Brief descriptions of these four types of correlated data follow.

1.1.1 Repeated measurements

Repeated measurements may be defined as data where a response variable for each experimental unit or subject is observed on multiple occasions and possibly under different experimental conditions. Within this context, the overall correlation structure among repeated measurements should be flexible enough to reflect the serial nature in which measurements are collected per subject while also accounting for possible correlation associated with subject-specific random effects. We have included longitudinal data under the heading of repeated measurements for several reasons:

Longitudinal data may be thought of as repeated measurements where the underlying metameter for the occasions at which measurements are taken is time. In this setting, interest generally centers on modeling and comparing trends over time. Consequently,

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

longitudinal data will generally exhibit some form of serial correlation much as one would expect with time-series data. In addition, one may also model correlation induced by a random-effects structure that allows for within- and between-subject variability. *Repeated measurements* are more general than longitudinal data in that the underlying metameter may be time or it may be a set of experimental conditions (e.g., different dose levels of a drug). Within this broader context, one may be interested in modeling and comparing trends over the range of experimental conditions (e.g., dose-response curves) or simply comparing mean values across different experimental conditions (e.g., a repeated measures analysis of variance).

Some typical settings where one is likely to encounter repeated measurements and longitudinal data include:

- Pharmacokinetic studies
  Studies where the plasma concentration of a drug is measured at several time points for each subject with an objective of estimating various population pharmacokinetic parameters.
- Econometrics
  Panel data entailing both cross-sectional and time-series data together in a two-dimensional array.
- Crossover studies
  Bioavailability studies, for example, routinely employ two-period, two-treatment crossover designs (e.g., AB | BA) where each subject receives each treatment on each of two occasions.
- Growth curve studies
  – Pediatric studies examining the growth pattern of children.
  – Agricultural studies examining the growth pattern of plants.

To illustrate, we consider the orange tree growth curve data presented in Draper and Smith (1981, p. 524) and analyzed by Lindstrom and Bates (1990). The data consists of the trunk circumference (millimeters) measured over 7 different time points on each of five orange trees. As shown in Figure 1.1, growth patterns exhibit a trend of ever increasing variability over time; a pattern reflective of a random-effects structure.

While we have lumped longitudinal data together with repeated measurements, it should be noted that event-times data, i.e., data representing time to some event, may also be classified as longitudinal even though the event time may be a single outcome measure such as patient survival. We will examine the analysis of event times data within the context of joint modeling of event times with repeated measurements/longitudinal data.

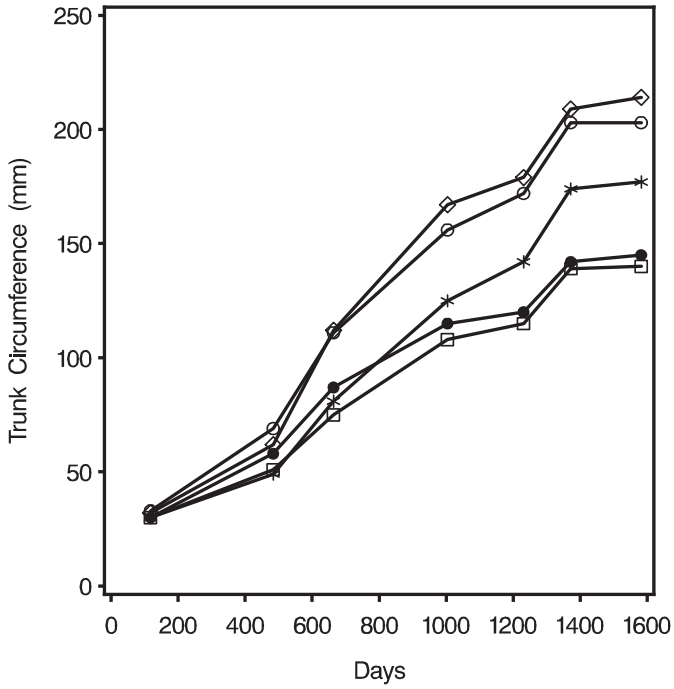
1.1.2 Clustered data

Clustered dependent data occur when observations are grouped in some natural fashion into clusters resulting in within-cluster data that tend to be correlated. Correlation induced by clustering is more often than not accounted for through the specification of a random-effects model in which cluster-specific random effects are used to differentiate within- and between-cluster variability. In some instances, there may be more than one level of clustering resulting in specification of a multi-level random-effects structure. Examples of clustered data include:

- Paired data
  – Studies on twins where each pair serves as a natural cluster.
  – Ophthalmology studies where a pair of eyes serves as a cluster.

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

Figure 1.1 Orange tree growth data



- Familial or teratologic studies
  - Studies on members of a litter of animals (e.g., toxicology studies).
  - Epidemiology studies of cancer where families serve as clusters.
- Agricultural studies
  Studies in which different plots of land serve as clusters and measurements within a plot are homogeneous.

1.1.3 Spatially correlated data

Spatially correlated data occur when observations include both a response variable and a location vector associated with that response variable. The location vector describes the position or location at which the measured response was obtained. The proximity of measured responses with one another determines the extent to which they are correlated. Lattice data, where measurements are linked to discrete regions (e.g., townships, counties, etc.) rather than some continuous coordinate system, are also considered as spatially correlated and are usually obtained from administrative data sources like census data, socio-economical data, and health data. Examples of spatially correlated data include:

- Geostatistical data
  Forestry and agronomy studies where sampling from specified (fixed) locations is used to draw inference over an entire region accounting for spatial dependencies.
  Mineral and petroleum exploration studies where the objective is more likely to be predictive in nature. Here one utilizes spatial variability patterns to help improve one's ability to predict resources in unmeasured locations.
- Epidemiological studies
  Studies aimed at describing the incidence and prevalence of a particular disease often use spatial correlation models in an attempt to smooth out region-specific counts so as to better assess potential environmental determinants and spatial patterns associated with the disease.

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

- Image analysis
  Image segmentation studies where the goal is to extract information about a particular region of interest from a given image. For example, in the field of medicine, image segmentation may be required to identify tissue regions that have been stained versus not stained. In these settings, modeling spatial correlation associated with a lattice array of pixel locations can help improve digital image analysis.

1.1.4 Multivariate data

Historically, the concept of correlation has been closely linked with methods for analyzing multivariate data wherein two or more response variables are measured per experimental unit or individual. Such methods include multivariate analysis of variance, cluster analysis, discriminant analysis, principal components analysis, canonical correlation analysis, etc. This book does not cover those topics but instead considers applications requiring the analysis of multivariate repeated measurements or the joint modeling of repeated measurements and one or more outcome measures that are measured only once. Examples we consider include

- Multivariate repeated measurements
  Any study where one has two or more outcome variables measured repeatedly over time.
- Joint modeling of repeated measurements and event-times data
  Studies where the primary goal is to draw inference on serial trends associated with a set of repeated measurements while accounting for possible informative censoring due to dropout.
  Studies where the primary goal is to draw joint inference on patient outcomes (e.g., patient survival) and any serial trends one might observe in a potential surrogate marker of patient outcome.

Of course one can easily imagine applications that involve two or more types of correlated data. For example, a longitudinal study may well entail both repeated measurements taken at the individual level and clustered data where groups of individuals form clusters according to some pre-determined criteria (e.g., a group randomized trial). Spatio-temporal data such as found in the mapping of disease rates over time is another example of combining two types of correlated data, namely spatially correlated data with serially correlated longitudinal data.

The majority of applications in this book deal with the analysis of repeated measurements, longitudinal data, clustered data, and to a lesser extent, spatial data. A more thorough treatment and illustration of applications involving the analysis of spatially correlated data and panel data, for example, may found in other texts including Cressie (1993), Littell et. al. (2006), Hsiao (2003), Frees (2004) and Mátyás and Sevestre (2008). We will also examine applications that require modeling multivariate repeated measurements in which two or more response variables are measured on the same experimental unit or individual on multiple occasions. As the focus of this book is on applications requiring the use of generalized linear and nonlinear models, examples will include methods for analyzing continuous, discrete and ordinal data including logistic regression for binary data, Poisson regression for counts, and nonlinear regression for normally distributed data.

1.2 Explanatory variables

When analyzing correlated data, one need first consider the kind of study from which the data were obtained. In this book, we consider data arising from two types of studies: 1) experimental studies, and 2) observational studies. Experimental studies are generally

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

interventional in nature in that two or more treatments are applied to experimental units with a goal of comparing the mean response across different treatments. Such studies may or may not entail randomization. For example, in a parallel group randomized placebo-controlled clinical trial, the experimental study may entail randomizing individuals into two groups; those who receive the placebo control versus those who receive an active ingredient. In contrast, in a simple pre-post study, a measured response is obtained on all individuals prior to receiving a planned intervention and then, following the intervention, a second response is measured. In this case, although no randomization is performed, the study is still experimental in that it does entail a planned intervention. Whether randomization is performed or not, additional explanatory variables are usually collected so as to 1) adjust for any residual confounding that may be present with or without randomization and 2) determine if interactions exist with the intervention.

In contrast to an experimental study, an observational study entails collecting data on available individuals from some target population. Such data would include any outcome measures of interest as well as any explanatory variables thought to be associated with the outcome measures.

In both kinds of studies, the set of explanatory variables, or covariates, used to model the mean response can be broadly classified into two categories:

- *Within-unit* factors or covariates (in longitudinal studies, these are often referred to as *time-dependent* covariates or *repeated measures* factors).
  For repeated measurements and longitudinal studies, examples include time itself, different dose levels applied to the same individual in a dose-response study, different treatment levels given to the same individual in a crossover study.
  For clustered data, examples include any covariates measured on individuals within a cluster.
- *Between-unit* factors or covariates (in longitudinal studies, these are often referred to as *time-independent* covariates)
  Examples include baseline characteristics in a longitudinal study (e.g., gender, race, baseline age), different treatment levels in a randomized prospective longitudinal study, different cluster-specific characteristics, etc.

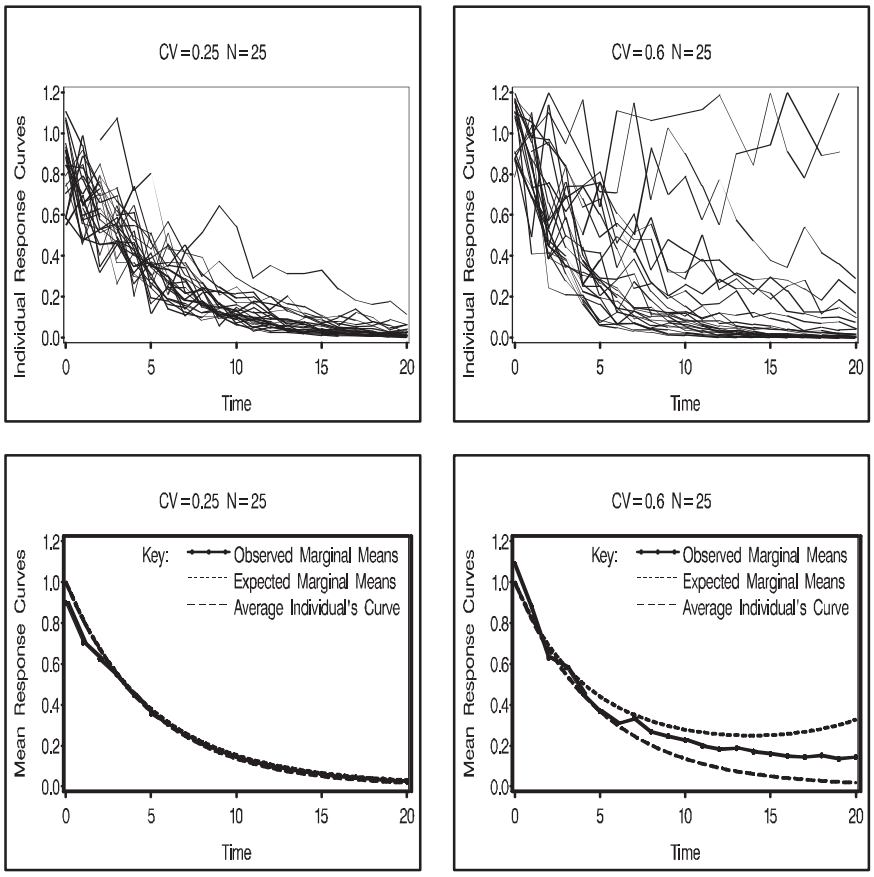
It is important to maintain the distinction between these two types of covariates for several reasons. One, it helps remind us that within-unit covariates model unit-specific trends while between-unit covariates model trends across units. Two, it may help in formulating an appropriate variance-covariance structure depending on the degree of heterogeneity between select groups of individuals or units. Finally, such distinctions are needed when designing a study. For example, sample size will be determined, in large part, on the primary goals of a study. When those goals focus on comparisons that involve within-unit covariates (either main effects or interactions), the number of experimental units needed will generally be less than when based on comparisons involving strictly between-unit comparisons.

1.3 Types of models

While there are several approaches to modeling correlated response data, we will confine ourselves to two basic approaches, namely the use of 1) marginal models and 2) mixed-effects models. With marginal models, the emphasis is on *population-averaged* (PA) inference where one focuses on the *marginal expectation of the responses*. Correlation is accounted for solely through specification of a *marginal* variance-covariance structure. The regression parameters of marginal models describe the population mean response and are most applicable in settings where the data are used to derive public policies. In contrast,

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

Figure 1.2 Individual and marginal mean responses under a simple negative exponential decay model with random decay rates



with a mixed-effects model, inference is more likely to be *subject-specific* (SS) or *cluster-specific* in scope with the focus centering on the *individual's* mean response. Here correlation is accounted for through specification of subject-specific random effects and possibly on an intra-subject covariance structure. Unlike marginal models, the fixed-effects regression parameters of mixed-effects models describe the average individual's response and are more informative when advising individuals of their expected outcomes. When a mixed-effects model is strictly linear in the random effects, the regression parameters will have both a population-averaged and subject-specific interpretation.

1.3.1 Marginal versus mixed-effects models

In choosing between marginal and mixed-effects models, one need carefully assess the type of inference needed for a particular application and weigh this against the complexities associated with running each type of model. For example, while a mixed-effects model may make perfect sense as far as its ability to describe heterogeneity and correlation, it may be extremely difficult to draw any population-based inference unless the model is strictly linear in the random effects. Moreover, population-based inference on, say, the marginal means may make little sense in applications where a mixed-effects model is assumed at the start. We illustrate this with the following example.

Shown in Figure 1.2 are the individual and marginal mean responses from a randomly generated sample of 25 subjects assuming a simple negative exponential decay model with

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

random decay rates. The model used to generate the data is given by:

$$y_{ij}|b_i = \exp\{-\beta_i t_{ij}\}[1+\epsilon_{ij}] \tag{1.1}$$
$$\beta_i = \beta + b_i \quad \text{with } \beta = 0.20,\ b_i \sim N(0,\sigma_b^2),\ \epsilon_{ij} \sim \text{iid}\, N(0,\sigma_\epsilon^2)$$

where y_{ij} is the response for the i^{th} subject $(i = 1,\ldots,n)$ on the j^{th} occasion, t_{ij} $(j = 1,\ldots,p)$, β_i is the i^{th} subject's decay rate, β is a population parameter decay rate, and b_i is a subject-specific random effect describing how far the i^{th} individual's decay rate deviates from the population parameter decay rate. Under this model, subject-specific (SS) inference targets the average individual's response curve while population-averaged (PA) inference targets the average response in the population. Specifically we have:

SS Inference - Average Subject's Response Curve:

$E_{y|b}[y_{ij}|b_i = 0] = E_{y|b}[y_{ij}|b_i = E_b(b_i)] = \exp\{-\beta t_{ij}\}$,

PA Inference - Population-Averaged Response Curve:

$E_y[y_{ij}] = E_b[E_{y|b}(y_{ij}|b_i)] = \exp\{-\beta t_{ij} + \frac{1}{2}\sigma_b^2 t_{ij}^2\}$.

Notice that for the average subject's response curve, expectation with respect to random effects is applied within the conditional argument (i.e., we are evaluating a conditional mean at the average random effect) while for the population-averaged response curve, expectation with respect to random effects is applied to the conditional means (i.e., we are evaluating the average response across subjects). Hence under model (1.1), the population-averaged (i.e., marginal) mean response depends on both the first and second moments of the subject-specific parameter, $\beta_i \sim N(\beta,\sigma_b^2)$.

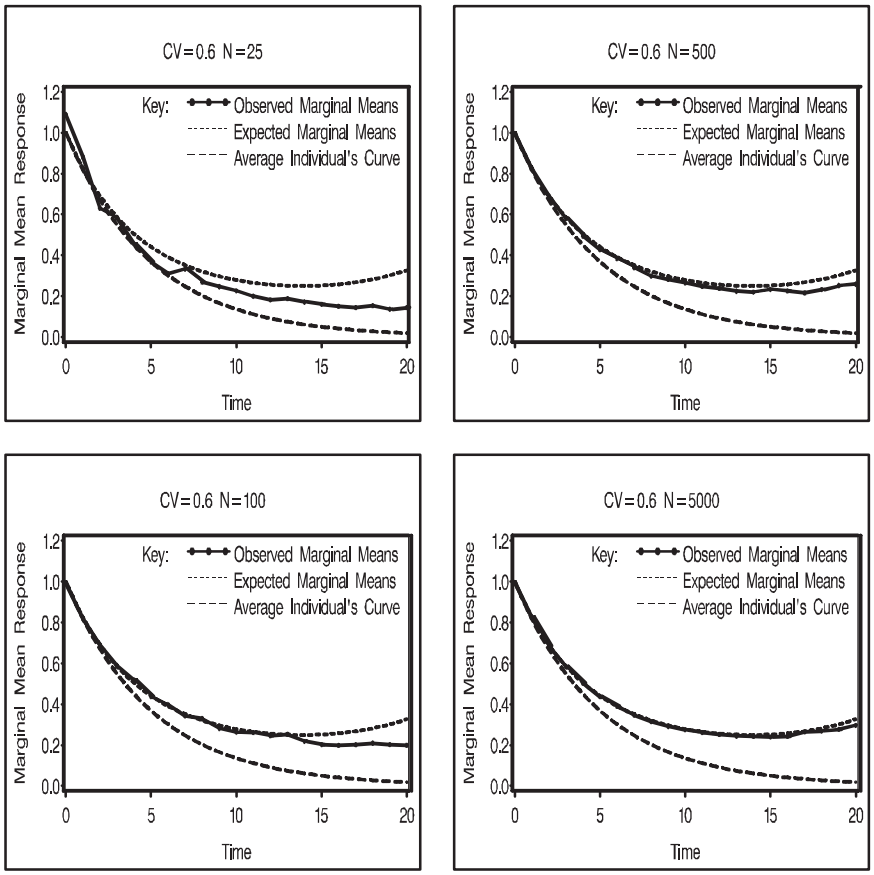
To contrast the SS response curves and PA response curves, we simulated data assuming 1) $\sigma_b/\beta = 0.25$ (i.e., a coefficient of variation, CV, of 25% with respect to β_i), and 2) $\sigma_b/\beta = 0.60$ (CV = 60%). As indicated in Figure 1.2, when the coefficient of variation of β_i increases, there is a clear divergence between the average subject's response curve and the mean response in the population. This reflects the fact that the marginal mean is a logarithmically convex function in t for all $t > \beta/\sigma_b^2$. It is also of interest to note that as the CV increases, the sample size required for the observed means to approximate the expected population means also increases (see Figure 1.3). One may question the validity of using simulated data where the response for some individuals actually increases over time despite the fact that the population parameter β depicts an overall decline over time (here, this will occur whenever the random effect $b_i < -0.20$ as $-\beta_i$ will then be positive). However, such phenomena do occur. For example, in section 1.4, we consider a study among patients with end-stage renal disease where their remaining kidney function, as measured by the glomerular filtration rate (GFR), tends to decline exponentially over time but for a few patients, their GFR actually increases as they regain their renal function.

1.3.2 Models in SAS

There are a variety of SAS procedures and macros available to users seeking to analyze correlated data. Depending on the type of data (discrete, ordinal, continuous, or a combination thereof), one can choose from one of four basic categories of models available in SAS: linear models, generalized linear models, nonlinear models and generalized nonlinear models. Within each basic category one can also choose to run a marginal model or a mixed-effects model resulting in the following eight classes of models available in SAS: 1) Linear Models (LM); 2) Linear Mixed-Effects Models (LME models); 3) Generalized Linear Models (GLIM); 4) Generalized Linear Mixed-Effects Models (GLME models); 5) Nonlinear Models (NLM); 6) Nonlinear Mixed-Effects Models (NLME models); 7) Generalized Nonlinear Models (GNLM); and 8) Generalized Nonlinear Mixed-Effects

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

Figure 1.3 Observed and expected marginal means for different sample sizes



Models (GNLME models). The models within any one class are determined through specification of the moments and possibly the distribution functions under which the data are generated. Moment-based specifications usually entail specifying unconditional (marginal) or conditional (mixed-effects) means and variances in terms of their dependence on covariates and/or random effects. Using the term "subject" to refer to an individual, subject, cluster or experimental unit, we adopt the following general notation which we use to describe marginal or conditional moments and/or likelihood functions.

Notation

$\boldsymbol{y}$ is a response vector for a given subject. Within a given context or unless otherwise noted, lower case lettering $\boldsymbol{y}$ (or y) will be used to denote either the underlying random vector (or variable) or its realization.
$\boldsymbol{\beta}$ is a vector of fixed-effects parameters associated with first-order moments (i.e., marginal or conditional means).
$\boldsymbol{b}$ is a vector of random-effects (possibly multi-level) which, unless otherwise indicated, is assumed to have a multivariate normal distribution, $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{\Psi})$, with variance-covariance matrix $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\boldsymbol{\theta}_b)$ that depends on a vector of between-subject variance-covariance parameters, say $\boldsymbol{\theta}_b$.
$\boldsymbol{X}$ is a matrix of within- and between-subject covariates linking the fixed-effects parameter vector $\boldsymbol{\beta}$ to the marginal or conditional mean.
$\boldsymbol{Z}$ is a matrix of within-subject covariates contained in $\boldsymbol{X}$ that directly link the random effects to the conditional mean.
$\boldsymbol{\mu}(\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{b}) = \boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{b}) = E(\boldsymbol{y}|\boldsymbol{b})$ is the conditional mean of $\boldsymbol{y}$ given random effects $\boldsymbol{b}$.

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

Table 1.1 Hierarchy of models in SAS according to mean structure and cumulative distribution function (CDF)

Marginal Models			Mixed-Effects Models		
Model	$\boldsymbol{\mu}(\boldsymbol{X},\boldsymbol{\beta})$	CDF	Model	$\mu(\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{Z},\boldsymbol{b})$	CDF
LM	$\boldsymbol{X\beta}$	Normal	LME	$\boldsymbol{X\beta}+\boldsymbol{Zb}$	Normal
GLIM	$g^{-1}(\boldsymbol{X\beta})$	General	GLME	$g^{-1}(\boldsymbol{X\beta}+\boldsymbol{Zb})$	General
NLM	$f(\boldsymbol{X},\boldsymbol{\beta})$	Normal	NLME	$f(\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{b})$	Normal
GNLM	$f(\boldsymbol{X},\boldsymbol{\beta})$	General	GNLME	$f(\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{b})$	General

$\boldsymbol{\Lambda}(\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{Z},\boldsymbol{b},\boldsymbol{\theta}_w)=\boldsymbol{\Lambda}(\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{\theta}_w)=Var(\boldsymbol{y}|\boldsymbol{b})$ is the conditional covariance matrix of $\boldsymbol{y}$ given random effects $\boldsymbol{b}$. This matrix may depend on an additional vector, $\boldsymbol{\theta}_w$, of within-subject variance-covariance parameters.
$\boldsymbol{\mu}(\boldsymbol{X},\boldsymbol{\beta})=\boldsymbol{\mu}(\boldsymbol{\beta})=E(\boldsymbol{y})$ is the marginal mean of $\boldsymbol{y}$ except for mixed models where the marginal mean $\boldsymbol{\mu}(\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{Z},\boldsymbol{\theta}_b)=E(\boldsymbol{y})$ may depend on $\boldsymbol{\theta}_b$ as well as $\boldsymbol{\beta}$ (e.g., see the PA mean response for model (1.1)).
$\boldsymbol{\Sigma}(\boldsymbol{\beta},\boldsymbol{\theta})=Var(\boldsymbol{y})$ is the marginal variance-covariance of $\boldsymbol{y}$ that depends on between- and/or within-subject covariance parameters, $\boldsymbol{\theta}=(\boldsymbol{\theta}_b,\boldsymbol{\theta}_w)$ and possibly on the fixed-effects regression parameters $\boldsymbol{\beta}$.

There is an inherent hierarchy to these models as suggested in Table 1.1. Specifically, linear models can be considered a special case of generalized linear models in that the latter allow for a broader class of distributions and a more general mean structure given by an inverse link function, $g^{-1}(\boldsymbol{X\beta})$, which is a monotonic invertible function linking the mean, $E(\boldsymbol{y})$, to a linear predictor, $\boldsymbol{X\beta}$, via the relationship $g(E(\boldsymbol{y}))=\boldsymbol{X\beta}$. Likewise Gaussian-based linear models are a special case of Gaussian-based nonlinear models in that the latter allow for a more general nonlinear mean structure, $f(\boldsymbol{X},\boldsymbol{\beta})$, rather than one that is strictly linear in the parameters of interest. Finally, generalized linear models are a special case of generalized nonlinear models in that the latter, in addition to allowing for a broader class of distributions, also allow for more general nonlinear mean structures of the form $f(\boldsymbol{X},\boldsymbol{\beta})$. That is, generalized nonlinear models do not require a mean structure that is an invertible monotonic function of a linear predictor as is the case for generalized linear models. In like fashion, within any row of Table 1.1, one can consider the marginal model as a special case of the mixed-effect model in that the former can be obtained by merely setting the random effects of the mixed-effects model to 0. We also note that since nonlinear mixed models do not require specification of a conditional linear predictor, $\boldsymbol{X\beta}+\boldsymbol{Zb}$, the conditional mean, $f(\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{b})$, may be specified without reference to $\boldsymbol{Z}$. This is because the fixed and random effects parameters, $\boldsymbol{\beta}$ and $\boldsymbol{b}$, will be linked to the appropriate covariates through specification of the function f and its relationship to a design matrix $\boldsymbol{X}$ that encompasses $\boldsymbol{Z}$.

The generalized nonlinear mixed-effects (GNLME) model is the most general model considered in that it combines the flexibility of a generalized linear model in terms of its ability to specify non-Gaussian distributions, and the flexibility of a nonlinear model in terms of its ability to specify more general mean structures. One might wonder, then, why SAS does not develop a single procedure based on the GNLME model rather than the various procedures currently available in SAS. The answer is simple. The various procedures specific to linear (MIXED, GENMOD and GLIMMIX) and generalized linear models (GENMOD, GLIMMIX) offer specific options and computational features that take full advantage of the inherent structure of the underlying model (e.g., the linearity of all parameters, both fixed and random, in the LME model, or the monotonic transformation that links the mean function to a linear predictor in a generalized linear model). Such flexibility is next to impossible to incorporate under NLME and GNLME models, both of

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

which can be fit to data using either the SAS procedure NLMIXED or the SAS macro %NLINMIX. A "road map" linking the SAS procedures and their key statements/options to these various models is presented in Table 1.2 of the summary section at the end of this chapter.

Finally, we shall assume throughout that the design matrices, $\boldsymbol{X}$ and $\boldsymbol{Z}$, are of full rank such that, where indicated, all matrices are invertible. In those cases where we are dealing with less than full rank design matrices and matrices of the form $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$, for example, are not of full rank, the expression $(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X})^{-1}$ will be understood to represent a generalized inverse of $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$.

1.3.3 Alternative approaches

Alternative approaches to modeling correlated response data include the use of conditional and/or transition models as well as hierarchical Bayesian models. Texts by Diggle, Liang and Zeger (1994) and Molenberghs and Verbeke (2005), for example, provide an excellent source of information on conditional and transition models for longitudinal data. Also, with the advent of Markov Chain Monte Carlo (MCMC) and other techniques for generating samples from Bayesian posterior distributions, interested practitioners can opt for a full Bayesian approach to modeling correlated data as exemplified in the texts by Carlin and Louis (2000) and Gelman et. al. (2004). A number of Bayesian capabilities were made available with the release of SAS 9.2 including the MCMC procedure. With PROC MCMC, users can fit a variety of Bayesian models using a general purpose MCMC simulation procedure.

1.4 Some examples

In this section, we present just a few examples illustrating the different types of response data, covariates, and models that will be discussed in more detail in later chapters.

Soybean Growth Data

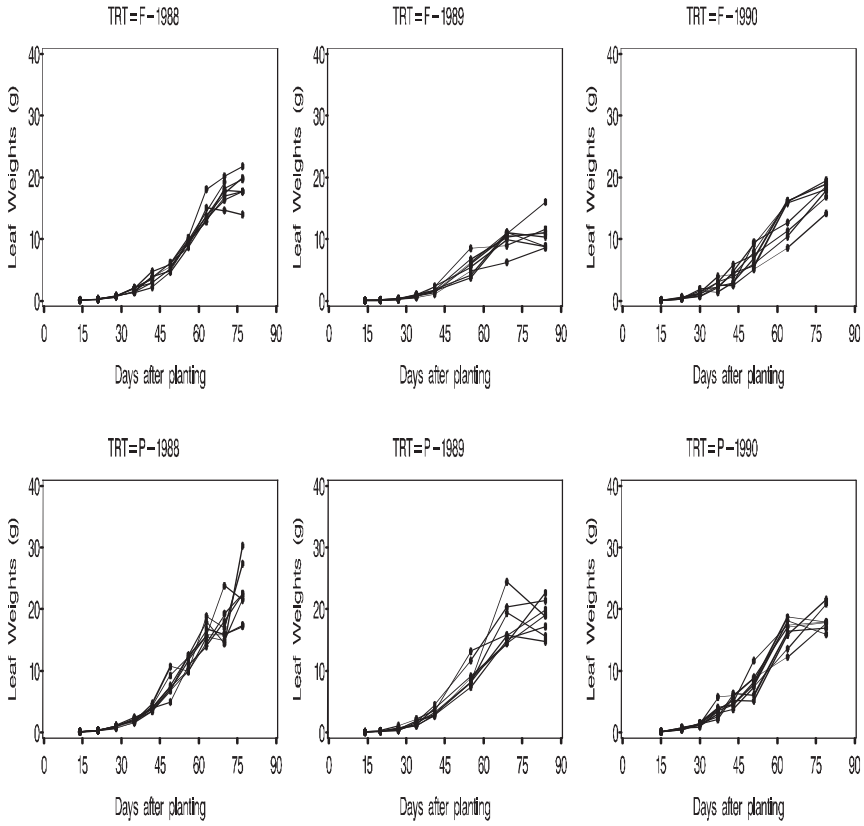
Davidian and Giltinan (1993, 1995) describe an experimental study in which the growth patterns of two genotypes of soybeans were to be compared. The essential features of the study are as follows:

- The experimental unit or cluster is a plot of land
- Plots were sampled 8-10 occasions (times) within a calendar year
- Six plants were randomly selected at each occasion and the average leaf weight per plant was calculated for a plot
- Response variable:
  – y_{ij} = average leaf weight (g) per plant for i^{th} plot on the j^{th} occasion (time) within a calendar year ($i = 1, \ldots, n = 48$ plots; 16 plots per each of the calendar years 1988, 1989 and 1990; $j = 1, \ldots, p_i$ with $p_i = 8$ to 10 measurements per calendar year)
- One within-unit covariate:
  – t_{ij} =days after planting for i^{th} plot on the j^{th} occasion
- Two between-unit covariates:
  – Genotype of Soybean (F=commercial, P=experimental) denoted by

$$a_{1i} = \begin{cases} 0, & \text{if commercial (F)} \\ 1, & \text{if experimental (P)} \end{cases}$$

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

Figure 1.4 Soybean growth data



– Calendar Year (1988, 1989, 1990) denoted by two indicator variables,

$$a_{2i} = \begin{cases} 0, & \text{if year is 1988 or 1990} \\ 1, & \text{if year is 1989} \end{cases}$$

$$a_{3i} = \begin{cases} 0, & \text{if year is 1988 or 1989} \\ 1, & \text{if year is 1990} \end{cases}$$

- Goal: compare the growth patterns of the two genotypes of soybean over the three growing seasons represented by calendar years 1988-1990.

This is an example of an experimental study involving clustered longitudinal data in which the response variable, y =average leaf weight per plant (g), is measured over time within plots (clusters) of land. In each of the three years, 1988, 1989 and 1990, 8 different plots of land were seeded with the genotype Forrest (F) representing a commercial strain of seeds and 8 different plots were seeded with genotype Plant Introduction #416937 (P), an experimental strain of seeds. During the growing season of each calendar year, six plants were randomly sampled from each plot on a weekly basis (starting roughly two weeks after planting) and the leaves from these plants were collected and weighed yielding an average leaf weight per plant, y, per plot. A plot of the individual profiles, shown in Figure 1.4, suggest a nonlinear growth pattern which Davidian and Giltinan modeled using a nonlinear

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

mixed-effects logistic growth curve model. One plausible form of this model might be

$$\begin{aligned} y_{ij} &= f(\boldsymbol{x}'_{ij}, \boldsymbol{\beta}, \boldsymbol{b}_i) + \epsilon_{ij} \\ &= f(\boldsymbol{x}'_{ij}, \boldsymbol{\beta}_i) + \epsilon_{ij} \\ &= \frac{\beta_{i1}}{1 + \exp\{\beta_{i3}(t_{ij} - \beta_{i2})\}} + \epsilon_{ij} \\ \boldsymbol{\beta}_i &= \begin{pmatrix} \beta_{i1} \\ \beta_{i2} \\ \beta_{i3} \end{pmatrix} = \begin{pmatrix} \beta_{01} + \beta_{11}a_{1i} + \beta_{21}a_{2i} + \beta_{31}a_{3i} \\ \beta_{02} + \beta_{12}a_{1i} + \beta_{22}a_{2i} + \beta_{32}a_{3i} \\ \beta_{03} + \beta_{13}a_{1i} + \beta_{23}a_{2i} + \beta_{33}a_{3i} \end{pmatrix} + \begin{pmatrix} b_{i1} \\ b_{i2} \\ b_{i3} \end{pmatrix} \end{aligned} \tag{1.2}$$

where y_{ij} is the average leaf weight per plant for the i^{th} plot on the j^{th} occasion, $f(\boldsymbol{x}'_{ij}, \boldsymbol{\beta}, \boldsymbol{b}_i) = f(\boldsymbol{x}'_{ij}, \boldsymbol{\beta}_i) = \beta_{i1}/\big[1 + \exp\{\beta_{i3}(t_{ij} - \beta_{i2})\}\big]$ is the conditional mean response for the i^{th} plot on the j^{th} occasion, $\boldsymbol{x}'_{ij} = \begin{pmatrix} 1 & t_{ij} & a_{1i} & a_{2i} & a_{3i} \end{pmatrix}$ is the vector of within- and between-cluster covariates associated with the population parameter vector, $\boldsymbol{\beta}' = \begin{pmatrix} \beta_{01} & \beta_{11} & \beta_{21} & \beta_{31} & \beta_{02} & \ldots & \beta_{23} & \beta_{33} \end{pmatrix}$ on the j^{th} occasion. The first two columns of $\boldsymbol{x}'_{ij}$, say $\boldsymbol{z}'_{ij} = \begin{pmatrix} 1 & t_{ij} \end{pmatrix}$, is the vector of within-cluster covariates associated with the cluster-specific random effects, $\boldsymbol{b}'_i = \begin{pmatrix} b_{i1} & b_{i2} & b_{i3} \end{pmatrix}$ on the j^{th} occasion, and ϵ_{ij} is the intra-cluster (within-plot or within-unit) error on the j^{th} occasion. The vector of random-effects are assumed to be iid $N(\mathbf{0}, \boldsymbol{\Psi})$ where $\boldsymbol{\Psi}$ is a 3×3 arbitrary positive definite covariance matrix. One should note that under this particular model, we are investigating the main effects of genotype and calendar year on the mean response over time.

The vector $\boldsymbol{\beta}'_i = \begin{pmatrix} \beta_{i1} & \beta_{i2} & \beta_{i3} \end{pmatrix}$ may be regarded as a cluster-specific parameter vector that uniquely describes the mean response for the i^{th} plot with $\beta_{i1} > 0$ representing the limiting growth value (asymptote), $\beta_{i2} > 0$ representing soybean "half-life" (i.e., the time at which the soybean reaches half its limiting growth value), and $\beta_{i3} < 0$ representing the growth rate. It may be written in terms of the linear random-effects model

$$\boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{b}_i$$

where

$$\boldsymbol{A}_i = \begin{pmatrix} 1 & a_{1i} & a_{2i} & a_{3i} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & a_{1i} & a_{2i} & a_{3i} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & a_{1i} & a_{2i} & a_{3i} \end{pmatrix}$$

is a between-cluster design matrix linking the between-cluster covariates, genotype and calendar year, to the population parameters $\boldsymbol{\beta}$ while $\boldsymbol{B}_i$ is an incidence matrix of 0's and 1's indicating which components of $\boldsymbol{\beta}_i$ are random and which are strictly functions of the fixed-effect covariates. In our current example, all three components of $\boldsymbol{\beta}_i$ are assumed random and hence $\boldsymbol{B}_i = \boldsymbol{I}_3$ is simply the identity matrix. Suppose, however, that the half-life parameter, β_{i2}, was assumed to be a function solely of the fixed-effects covariates. Then $\boldsymbol{B}_i$ would be given by

$$\boldsymbol{B}_i = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}$$

with $\boldsymbol{b}'_i = \begin{pmatrix} b_{i1} & b_{i3} \end{pmatrix}$. Note, too, that we can express the between-unit design matrix more conveniently as $\boldsymbol{A}_i = \boldsymbol{I}_3 \otimes \begin{pmatrix} 1 & a_{1i} & a_{2i} & a_{3i} \end{pmatrix}$ where $\otimes$ is the direct product operator (or Kronecker product) linking the dimension of $\boldsymbol{\beta}_i$ via the identity matrix $\boldsymbol{I}_3$ to the between-unit covariate vector, $\boldsymbol{a}'_i = \begin{pmatrix} 1 & a_{1i} & a_{2i} & a_{3i} \end{pmatrix}$. The $\otimes$ operator is a useful tool

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

which we will have recourse to use throughout the book and which is described more fully in Appendix A.

Finally, by assuming the intra-cluster errors ϵ_{ij} are iid $N(0, \sigma_w^2)$, model (1.2) may be classified as a nonlinear mixed-effects model (NLME) having conditional means expressed in terms of a nonlinear function, i.e., $E(y_{ij}|\boldsymbol{b}_i) = \mu(\boldsymbol{x}'_{ij}, \boldsymbol{\beta}, \boldsymbol{z}'_{ij}, \boldsymbol{b}_i) = f(\boldsymbol{x}'_{ij}, \boldsymbol{\beta}, \boldsymbol{b}_i)$, as represented using the general notation of Table 1.1. In this case, inference with respect to the population parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_b, \boldsymbol{\theta}_w) = (vech(\boldsymbol{\Psi}), \sigma_w^2)$ will be cluster-specific in scope in that the function f, when evaluated at $\boldsymbol{b}_i = \boldsymbol{0}$, describes what the average soybean growth pattern is for a "typical plot."

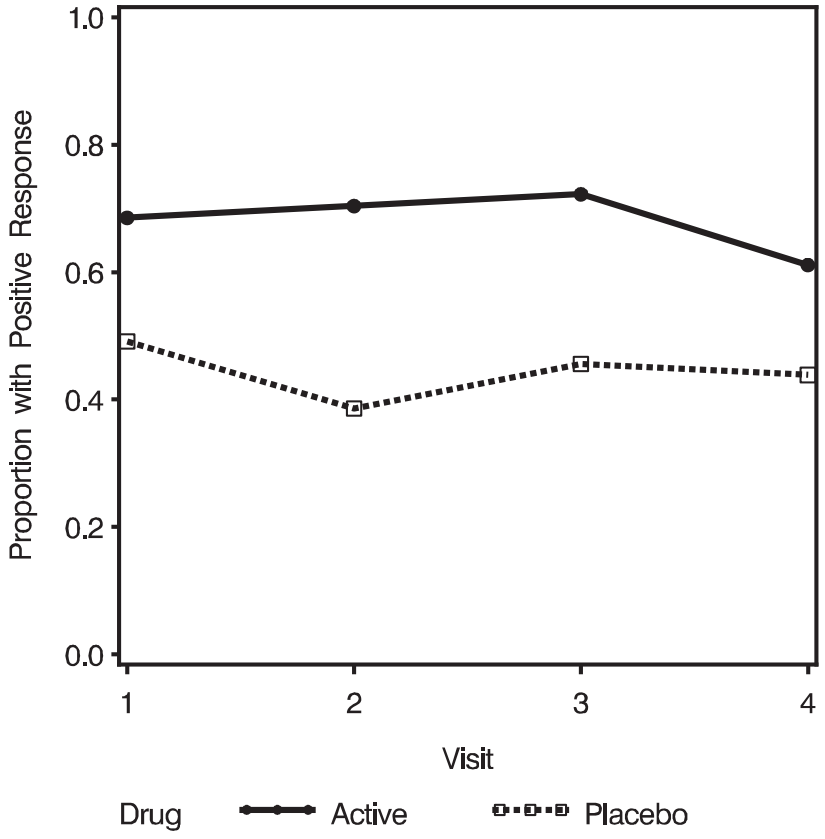
Respiratory Disorder Data

Koch et. al. (1990) and Stokes et. al. (2000) analyzed data from a multicenter randomized controlled trial for patients with a respiratory disorder. The trial was conducted in two centers in which patients were randomly assigned to one of two treatment groups: an active treatment group or a placebo control group (0 = placebo, 1 = active). The initial outcome variable of interest was patient status which is defined in terms of the ordinal categorical responses: 0 = terrible, 1 = poor, 2 = fair, 3 = good, 4 = excellent. This categorical response was obtained both at baseline and at each of four visits (visit 1, visit 2, visit 3, visit 4) during the course of treatment. Here we consider an analysis obtained by collapsing the data into a discrete binary outcome with the primary focus being a comparison of the average response of patients. The essential components of the study are listed below (the SAS variables are listed in parentheses).

- The experimental unit is a patient (identified by SAS variable ID)
- The initial outcome variable was patient status defined by the ordinal response: 0 = terrible, 1 = poor, 2 = fair, 3 = good, 4 = excellent
- Response variable (y):
  – The data were collapsed into a simple binary response as
  $$y_{ij} = \begin{cases} 0 = \text{negative response if (terrible, poor, or fair)} \\ 1 = \text{positive response if (good, excellent)} \end{cases}$$
  which is the i^{th} patient's response obtained on the j^{th} visit
  • One within-unit covariate:
  – Visit (1, 2, 3, or 4) defined here by the indictor variables (Visit)
  $$v_{ij} = \begin{cases} 1, & \text{if visit } j \\ 0, & \text{otherwise} \end{cases}$$
- Five between-unit covariates:
  – Treatment group (Treatment) defined as 'P' for placebo or 'A' for active.
  $$a_{1i} = \begin{cases} 0, & \text{if placebo} \\ 1, & \text{if active} \end{cases} \quad \text{(this indicator is labeled Drug0)}$$
  – Center (Center)
  $$a_{2i} = \begin{cases} 0, & \text{if center 1} \\ 1, & \text{if center 2} \end{cases} \quad \text{(this indicator is labeled Center0)}$$
  – Gender (Gender) defined as 0 = male, 1 = female
  $$a_{3i} = \begin{cases} 0, & \text{if male} \\ 1, & \text{if female} \end{cases} \quad \text{(this indicator is labeled Sex)}$$
  – Age at baseline (Age)
  a_{4i} =patient age

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

Figure 1.5 A plot of the proportion of positive responses by visit and drug for the respiratory disorder data



– Response at baseline (Baseline),

$$a_{5i} = \begin{cases} 0 = \text{negative}, \\ 1 = \text{positive} \end{cases} \quad \text{(this indicator is labeled y0)}$$

- Goal: determine if there is a treatment effect after adjusting for center, gender, age and baseline differences.

A plot of the proportion of positive responses at each visit according to treatment group is shown in Figure 1.5. In this example, previous authors (Koch et. al., 1977; Koch et. al., 1990) fit the data using several different marginal models resulting in a population-averaged approach to inference. One family of such models is the family of marginal generalized linear models (GLIM's) with working correlation structure. For example, under the assumption of a working independence structure across visits and assuming there is no visit effect, one might fit this data to a binary logistic regression model of the form

$$\begin{aligned} E(y_{ij}) &= \mu_{ij}(\boldsymbol{x}'_{ij}, \boldsymbol{\beta}) = g^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta}) \\ &= \frac{\exp\{\boldsymbol{x}'_{ij}\boldsymbol{\beta}\}}{1 + \exp\{\boldsymbol{x}'_{ij}\boldsymbol{\beta}\}} \\ \boldsymbol{x}'_{ij}\boldsymbol{\beta} &= \beta_0 v_{ij} + \beta_1 a_{1i} + \beta_2 a_{2i} + \beta_3 a_{3i} + \beta_4 a_{4i} + \beta_5 a_{5i} \\ Var(y_{ij}) &= \mu_{ij}(\boldsymbol{x}'_{ij}, \boldsymbol{\beta})[1 - \mu_{ij}(\boldsymbol{x}'_{ij}, \boldsymbol{\beta})] = \mu_{ij}(1 - \mu_{ij}) \end{aligned} \tag{1.3}$$

where $\mu_{ij} = \mu_{ij}(\boldsymbol{x}'_{ij}, \boldsymbol{\beta}) = \Pr(y_{ij} = 1 | \boldsymbol{x}_{ij})$ is the probability of a positive response on the j^{th} visit, $\boldsymbol{x}'_{ij} = \begin{pmatrix} v_{ij} & a_{1i} & a_{2i} & a_{3i} & a_{4i} & a_{5i} \end{pmatrix}$ is the design vector of within- and between-unit

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

covariates on the j^{th} visit and $\boldsymbol{\beta}' = \begin{pmatrix}\beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5\end{pmatrix}$ is the parameter vector associated with the mean response. Here, we do not model a visit effect but rather assume a common visit effect that is reflected in the overall intercept parameter, β_0 (i.e., since we always have $v_{ij} = 1$ on the j^{th} visit, we can simply replace $\beta_0 v_{ij}$ with β_0). In matrix notation, model (1.3) may be written as

$$E(\boldsymbol{y}_i) = \mu_i(\boldsymbol{X}_i, \boldsymbol{\beta}) = g^{-1}(\boldsymbol{X}_i\boldsymbol{\beta})$$
$$Var(\boldsymbol{y}_i) = \boldsymbol{\Sigma}_i(\boldsymbol{\beta}) = \boldsymbol{H}_i(\boldsymbol{\mu}_i)^{1/2}\boldsymbol{I}_4\boldsymbol{H}_i(\boldsymbol{\mu}_i)^{1/2}$$

where

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{pmatrix}, \; \boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}'_{i1} \\ \boldsymbol{x}'_{i2} \\ \boldsymbol{x}'_{i3} \\ \boldsymbol{x}'_{i4} \end{pmatrix},$$

and $\boldsymbol{H}_i(\boldsymbol{\mu}_i)$ is the 4×4 diagonal variance matrix with $Var(y_{ij}) = \mu_{ij}(1 - \mu_{ij})$ as the j^{th} diagonal element and $\boldsymbol{H}_i(\boldsymbol{\mu}_i)^{1/2}$ is the square root of $\boldsymbol{H}_i(\boldsymbol{\mu}_i)$ which, for arbitrary positive definite matrices, may be obtained via the Cholesky decomposition. To accommodate possible correlation among binary responses taken on the same subject across visits, we can use a generalized estimating equation (GEE) approach in which robust standard errors are computed using the so-called empirical sandwich estimator (e.g., see §4.2 of Chapter 4). An alternative GEE approach might assume an overdispersion parameter ϕ and "working" correlation matrix $\boldsymbol{R}_i(\boldsymbol{\alpha})$ such that

$$Var(\boldsymbol{y}_i) = \boldsymbol{\Sigma}_i(\boldsymbol{\beta}, \boldsymbol{\theta}) = \phi\boldsymbol{H}_i(\boldsymbol{\mu}_i)^{1/2}\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{H}_i(\boldsymbol{\mu}_i)^{1/2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \phi)$. Commonly used working correlation structures include working independence [i.e., $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \boldsymbol{I}_4$ as above], compound symmetry and first-order autoregressive. Alternatively, one can extend the GLIM in (1.3) to a generalized linear mixed-effects model (GLME) by simply adding a random intercept term, b_i, to the linear predictor, i.e.,

$$\boldsymbol{x}'_{ij}\boldsymbol{\beta} + b_i = \beta_0 v_{ij} + \beta_1 a_{1i} + \beta_2 a_{2i} + \beta_3 a_{3i} + \beta_4 a_{4i} + \beta_5 a_{5i} + b_i.$$

This will induce a correlation structure among the repeated binary responses via the random intercepts shared by patients across their visits.

In Chapters 4-5, we shall consider analyses based on both marginal and mixed-effects logistic regression thereby allowing one to contrast PA versus SS inference. With respect to marginal logistic regression, we will present results from a first-order and second-order generalized estimating equation approach.

Epileptic Seizure Data

Thall and Vail (1990) and Breslow and Clayton (1993) used Poisson regression to analyze epileptic seizure data from a randomized controlled trial designed to compare the effectiveness of progabide versus placebo to reduce the number of partial seizures occurring over time. The key attributes of the study are summarized below (SAS variables denoted in parentheses).

- The experimental unit is a patient (ID)
- The data consists of the number of partial seizures occurring over a two week period on each of four successive visits made by patients receiving one of two treatments (progabide, placebo).

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

- Response variable (y):
  – y_{ij} =number of partial seizures in a two-week interval for the i^{th} patient as recorded on the j^{th} visit
- One within-unit covariate:
  – Visit (1, 2, 3, or 4) defined here by the indictor variables (Visit)
  $$v_{ij} = \begin{cases} 1, & \text{if visit } j \\ 0, & \text{otherwise} \end{cases}$$
- Three between-unit covariates:
  – Treatment group (Trt)
  $$a_{1i} = \begin{cases} 0, & \text{if placebo} \\ 1, & \text{if progabide} \end{cases}$$
  – Age at baseline (Age)
  a_{2i} =patient age
  – Baseline seizure counts (bline) normalized to a two week period
  $a_{3i} = \frac{1}{4} \times$baseline seizure counts over an 8 week period (y0)
- Goal: determine if progabide is effective in reducing the number of seizures after adjustment for relevant baseline covariates

In Chapters 4-5, we will consider several different models for analyzing this data using both a marginal and mixed-effects approach. For example, we consider several mixed-effects models in which the count data (y_{ij} =number of seizures per 2-week interval) were fitted to a mixed-effects log-linear model with conditional means of the form

$$\begin{aligned} E(y_{ij}|b_i) =& \mu_{ij}(\boldsymbol{x}'_{ij}, \boldsymbol{\beta}, z_{ij}, b_i) = g^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + z_{ij}b_i) \\ =& \exp\{\boldsymbol{x}'_{ij}\boldsymbol{\beta} + z_{ij}b_i\} \\ \boldsymbol{x}'_{ij}\boldsymbol{\beta} + z_{ij}b_i =& \beta_0 + \beta_1 a_{1i} + \beta_2 \log(a_{2i}) + \beta_3 \log(a_{3i}) + \\ & \beta_4 a_{1i} \log(a_{3i}) + \beta_5 v_{i4} + \log(2) + b_i \end{aligned} \tag{1.4}$$

where $\mu_{ij} = \mu_{ij}(\boldsymbol{x}'_{ij}, \boldsymbol{\beta}, z_{ij}, b_i)$ is the average number of seizures per two week period on the j^{th} visit, $\boldsymbol{x}'_{ij} = \begin{pmatrix} 1 & a_{1i} & \log(a_{2i}) & \log(a_{3i}) & a_{1i}\log(a_{3i}) & v_{i4} \end{pmatrix}$ is the design vector of within- and between-unit covariates on the j^{th} visit, $z_{ij} \equiv 1$ for each j, b_i is a subject-specifc random intercept, and $\boldsymbol{\beta}' = \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \end{pmatrix}$ is the parameter vector associated with the mean response. Following Thall and Vail (1990) and Breslow and Clayton (1993), this model assumes that only visit 4 has an effect on seizure counts. The term, log(2), is an offset that is included to reflect that the mean count is over a two-week period. In matrix notation, the conditional means across all four visits for a given subject may be written as

$$E(\boldsymbol{y}_i|b_i) = \boldsymbol{\mu}_i(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{Z}_i, b_i) = \boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{b}_i) = g^{-1}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i b_i)$$

where

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{pmatrix}, \; \boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}'_{i1} \\ \boldsymbol{x}'_{i2} \\ \boldsymbol{x}'_{i3} \\ \boldsymbol{x}'_{i4} \end{pmatrix}, \; \boldsymbol{Z}_i = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

One could consider fitting this model assuming any one of three conditional variance structures, $Var(y_{ij}|b_i) = \boldsymbol{\Lambda}_i(\boldsymbol{x}_{ij}, \boldsymbol{\beta}, \boldsymbol{z}_{ij}, b_i, \boldsymbol{\theta}_w) = \boldsymbol{\Lambda}_i(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\theta}_w)$,

Case 1: $Var(y_{ij}|b_i) = \mu_{ij}$ (standard Poisson variation)
Case 2: $Var(y_{ij}|b_i) = \phi\mu_{ij}$ (extra-Poisson variation)
Case 3: $Var(y_{ij}|b_i) = \mu_{ij}(1+\alpha\mu_{ij})$ (negative binomial variation).

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

In the first case, $\boldsymbol{\Lambda}_i(\boldsymbol{\beta}, b_i, \boldsymbol{\theta}_w) = \mu_{ij}(\boldsymbol{\beta}, b_i)$ and there is no $\boldsymbol{\theta}_w$ parameter. In the second case, we allow for conditional overdispersion in the form of $\boldsymbol{\Lambda}_i(\boldsymbol{\beta}, b_i, \boldsymbol{\theta}_w) = \phi\mu_{ij}(\boldsymbol{\beta}, b_i)$ where $\boldsymbol{\theta}_w = \phi$ while in the third case, over-dispersion in the form of a negative binomial model with $\boldsymbol{\Lambda}_i(\boldsymbol{\beta}, b_i, \boldsymbol{\theta}_w) = \mu_{ij}(\boldsymbol{\beta}, b_i)(1 + \alpha\mu_{ij}(\boldsymbol{\beta}, b_i))$ is considered with $\boldsymbol{\theta}_w = \alpha$. The conditional negative binomial model coincides with a conditional gamma-Poisson model which is obtained by assuming the conditional rates within each two-week interval are further distributed conditionally as a gamma random variable. This allows for a specific form of conditional overdispersion which may or may not make much sense in this setting. Assuming $b_i \sim$iid $N(0, \sigma_b^2)$, all three cases result in models that belong to the class of GLME models. However, the models in the first and third cases are based on a well-defined conditional distribution (Poisson in case 1 and negative binomial in case 3) which allows one to estimate the model parameters using maximum likelihood estimation. Under this same mixed-effects setting, other covariance structures could also be considered some of which when combined with the assumption of a conditional Poisson distribution yield a GNLME model generally not considered in most texts.

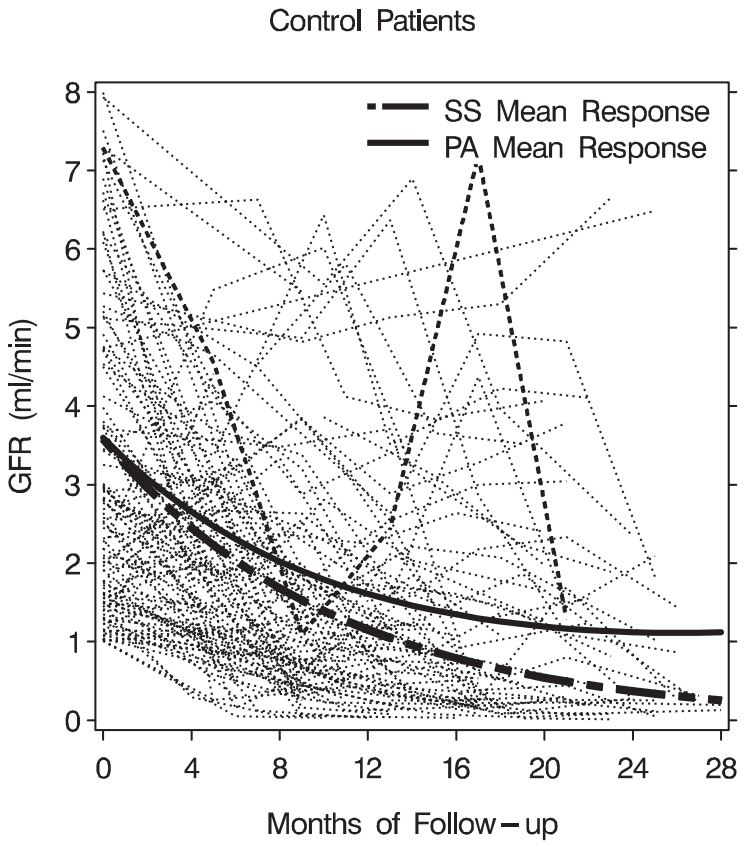
ADEMEX Data

Paniagua et. al. (2002) summarized results of the ADEMEX trial, a randomized multi-center trial of 965 Mexican patients designed to compare patient outcomes (e.g., survival, hospitalization, quality of life, etc.) among end-stage renal disease patients randomized to one of two dose levels of continuous ambulatory peritoneal dialysis (CAPD). While patient survival was the primary endpoint, there were a number of secondary and exploratory endpoints investigated as well. One such endpoint was the estimation and comparison of the decline in the glomerular filtration rate (GFR) among incident patients randomized to the standard versus high dose of dialysis. There were several challenges with this objective as described below. The essential features of this example are as follows:

- The experimental unit is an incident patient
- Response variables:
  – y_{ij} =glomerular filtration rate (GFR) of the kidney (ml/min) for i^{th} subject on the j^{th} occasion
  – T_i =patient survival time in months
- One within-unit covariate:
  – t_{ij} =months after randomization
- Six between-unit covariates:
  – Treatment group (standard dose, high dose)
  $$a_{1i} = \begin{cases} 0, & \text{if control = standard dose} \\ 1, & \text{if intervention = high dose} \end{cases}$$
  – Gender
  $$a_{2i} = \begin{cases} 0, & \text{if male} \\ 1, & \text{if female} \end{cases}$$
  – Age at baseline
  a_{3i} =patient age
  – Presence or absence of diabetes at baseline
  $$a_{4i} = \begin{cases} 0, & \text{if non-diabetic} \\ 1, & \text{if diabetic} \end{cases}$$
  – Baseline value of albumin
  a_{5i} =serum albumin (g/dL)
  - Baseline value of normalized protein nitrogen appearance (nPNA)
  a_{6i} =nPNA (g/kg/day)

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

Figure 1.6 SS and PA mean GFR profiles among control patients randomized to the standard dose of dialysis.



- Goal: Estimate the rate of decline in GFR and assess whether this rate differentially affects patient survival according to dose of dialysis
- Issues:
  1. Past studies have linked low GFR with increased mortality
  2. The analysis requires one to jointly model GFR and patient survival in order to determine if a) the rate of decline in GFR is associated with survival and b) if the rate of decline is affected by the dose of dialysis

As shown in Figure 1.6, the decline in GFR appears to occur more rapidly early on and then gradually reaches a value of 0 provided the patient lives long enough (i.e., the patient becomes completely anuric). Such data might be reasonably fit assuming a nonlinear exponential decay model with a random intercept and random decay rate. One such model is given by

$$\begin{aligned} y_{ij} &= f(\boldsymbol{x}'_{ij}, \boldsymbol{\beta}, \boldsymbol{b}_i) + \epsilon_{ij} \\ &= f(\boldsymbol{x}'_{ij}, \boldsymbol{\beta}_i) + \epsilon_{ij} \\ &= \beta_{i1} \exp(-\beta_{i2} t_{ij}) + \epsilon_{ij} \\ \boldsymbol{\beta}_i &= \begin{pmatrix} \beta_{i1} \\ \beta_{i2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}'_i \boldsymbol{\beta}_1 \\ \boldsymbol{a}'_i \boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} = \boldsymbol{A}_i \boldsymbol{\beta} + \boldsymbol{b}_i \end{aligned} \tag{1.5}$$

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

where $\boldsymbol{a}_i' = \begin{pmatrix} 1 & a_{1i} & a_{2i} & a_{3i} & a_{4i} & a_{5i} & a_{6i} \end{pmatrix}$ is the between-subject design vector, $\boldsymbol{x}_{ij}' = (t_{ij}, \boldsymbol{a}_i')$ is the vector of within- and between-subject covariates on the j^{th} occasion,

$$\begin{aligned} \boldsymbol{a}_i'\boldsymbol{\beta}_1 + b_{i1} = \ & \beta_{01} + \beta_{11}a_{1i} + \beta_{21}a_{2i} + \beta_{31}a_{3i} + \beta_{41}a_{4i} + \beta_{51}a_{5i} + \beta_{61}a_{6i} + b_{i1} \\ \boldsymbol{a}_i'\boldsymbol{\beta}_2 + b_{i2} = \ & \beta_{02} + \beta_{12}a_{1i} + \beta_{22}a_{2i} + \beta_{32}a_{3i} + \beta_{42}a_{4i} + \beta_{52}a_{5i} + \beta_{62}a_{6i} + b_{i2} \end{aligned}$$

are linear predictors for the intercept (β_{i1}) and decay rate (β_{i2}) parameters, respectively, $\boldsymbol{b}_i' = \begin{pmatrix} b_{i1} & b_{i2} \end{pmatrix}$ are the random intercept and decay rate effects, and $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ is the corresponding population parameter vector with $\boldsymbol{\beta}_1' = \begin{pmatrix} \beta_{01} & \beta_{11} & \beta_{21} & \beta_{31} & \beta_{41} & \beta_{51} & \beta_{61} \end{pmatrix}$ denoting the population intercept parameters and $\boldsymbol{\beta}_2' = \begin{pmatrix} \beta_{02} & \beta_{12} & \beta_{22} & \beta_{32} & \beta_{42} & \beta_{52} & \beta_{62} \end{pmatrix}$ the population decay rate parameters. Here $\boldsymbol{A}_i = \boldsymbol{I}_2 \otimes \boldsymbol{a}_i'$ is the between-subject design matrix linking the covariates $\boldsymbol{a}_i$ to $\boldsymbol{\beta}$ while $\epsilon_{ij} \sim$iid $N(0, \sigma_w^2)$ independent of $\boldsymbol{b}_i$.

Figure 1.6 reflects a reduced version of this model in which the subject-specific intercept and decay rate parameters are modeled as a simple linear function of the treatment group only, i.e.,

$$\begin{aligned} \beta_{i1} &= \beta_{01} + \beta_{11}a_{1i} + b_{i1} \\ \beta_{i2} &= \beta_{02} + \beta_{12}a_{1i} + b_{i2}. \end{aligned}$$

What is shown in Figure 1.6 is the estimated marginal mean profile (i.e., PA mean response) for the patients randomized to the control group. This PA mean response is obtained by averaging over the individual predicted response curves while the average patient's mean profile (i.e., the SS mean response) is obtained by simply plotting the predicted response curve achieved when one sets the random effects b_{i1} and b_{i2} equal to 0. As indicated previously (page 7), this example illustrates how one can have a random-effects exponential decay model that can predict certain individuals as having an increasing response over time. In this study, for example, two patients (one from each group) had a return of renal function while others showed either a modest rise or no change in renal function. The primary challenge here is to jointly model the decline in GFR and patient survival using a generalized nonlinear mixed-effects model that allows one to account for correlation in GFR values over time as well as determine if there is any association between the rate of decline in GFR over time and patient survival.

1.5 Summary features

We summarize here a number of features associated with the analysis of correlated data. First, since the response variables exhibit some degree of dependency as measured by correlation among the responses, most analyses may be classified as being essentially multivariate in nature. With repeated measurements and clustered data, for example, the analysis requires combining cross-sectional (between-cluster, between-subject, inter-subject) methods with time-series (within-cluster, within-subject, intra-subject) methods.

Second, the type of model one uses, marginal versus mixed, determines and/or limits the type of correlation structure one can model. In marginal models, correlation is accounted for by directly specifying an intra-subject or intra-cluster covariance structure. In mixed-effects models, a type of intraclass correlation structure is introduced through specification of subject-specific random effects. Specifically, intraclass correlation occurs as a result of having random effect variance components that are shared across measurements within subjects. Along with specifying what type of model, marginal or mixed, is needed for inferential purposes, one must also select what class of models is most appropriate based on the type of response variable being measured (e.g., continuous, ordinal, count or nominal) and its underlying mean structure. By specifying both the class of models and

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.

Table 1.2 A summary of the different classes of models and the type of model within a class that are available for the analysis of correlated response data. [1.] The SAS macro %NLINMIX iteratively calls the MIXED procedure when fitting nonlinear models. [2.] NLMIXED can be adpated to analyze marginal correlation structures (see §4.4, §4.5.2, §4.5.3, §4.5.4, §5.3.3).

	Types of	SAS	Type of Model	
Class of Models	**Data**	**PROC**	**Marginal[2.]**	**Mixed**
Linear	Continuous	MIXED	REPEATED	RANDOM
		GLIMMIX	RANDOM/RSIDE	RANDOM
Generalized Linear	Continuous	GENMOD	REPEATED	—
	Ordinal	GLIMMIX	RANDOM/RSIDE	RANDOM
	Count	NLMIXED	—	RANDOM
	Binary			
Generalized Nonlinear	Continuous	NLMIXED	—	RANDOM
	Ordinal	%NLINMIX[1.]	REPEATED	RANDOM
	Count			
	Binary			

type of model within a class, an appropriate SAS procedure can then be used to analyze the data. Summarized in Table 1.2 are the three major classes of models used to analyze correlated response data and the SAS procedure(s) and corresponding procedural statements/options for conducting such an analysis.

Another feature worth noting is that studies involving clustered data or repeated measurements generally lead to efficient within-unit comparisons but inefficient between-unit comparisons. This is evident with split-plot designs where we have efficient split-plot comparisons but inefficient whole plot comparisons. In summary, some of the features we have discussed in this chapter include:

- Repeated measurements, clustered data and spatial data are essentially multivariate in nature due to the correlated outcome measures
- Analysis of repeated measurements and longitudinal data often requires combining cross-sectional methods with time-series methods
- The analysis of correlated data, especially that of repeated measurements, clustered data and spatial data can be based either on marginal or mixed-effects models. Parameters from these two types of models generally differ in that
  1) Marginal models target population-averaged (PA) inference
  2) Mixed-effects models target subject-specific (SS) inference
- Marginal models can accommodate correlation via
  1) Direct specification of an intra-subject correlation structure
- Mixed-effects models can accommodate correlation via
  1) Direct specification of an intra-subject correlation structure
  2) Intraclass correlation resulting from random-effects variance components that are shared across observations within subjects
- The family of models available within SAS include linear, generalized linear, nonlinear, and generalized nonlinear models
- Repeated measurements and clustered data lead to efficient within-unit comparisons (including interactions of within-unit and between-unit covariates) but inefficient between-unit comparisons.

Vonesh, Edward F. Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore.