Introduction

1.1 Types of Models That Produce Data

Data sets presented in this book come from three types of sources: (1) designed experiments, (2) sample surveys, and (3) observational studies. Virtually all data sets are produced by one of these three sources.

In designed experiments, some form of treatment is applied to experimental units and responses are observed. For example, a researcher might want to compare two or more drug formulations to control high blood pressure. In a human clinical trial, the experimental units are volunteer patients who meet the criteria for participating in the study. The various drug formulations are randomly assigned to patients and their responses are subsequently observed and compared. In sample surveys, data are collected according to a plan, called a survey design, but treatments are

not applied to units. Instead, the units, typically people, already possess certain attributes such as age or occupation. It is often of interest to measure the effect of the attributes on, or their association with, other attributes. In observational studies, data are collected on units that are available, rather than on units chosen according to a plan. An example is a study at a veterinary clinic in which dogs entering the clinic are diagnosed according to their skin condition and blood samples are drawn for measurement of trace elements.

The objectives of a project, the types of resources that are available, and the constraints on what kind of data collection is possible all dictate your choice of whether to run a designed experiment, a sample survey, or an observational study. Even though the three have striking differences in the way they are carried out, they all have common features leading to a common terminology. For example, the terms **factor**, **level**, and **effect** are used alike in design experiments, sample surveys, and observational studies. In designed experiments, the treatment condition under study (e.g., from examples we decide to use) is the *factor* and the specific treatments are the *levels*. In the observational study, the dogs' diagnosis is the factor and the specific skin conditions are the levels. In all three types of studies, each level has an *effect*; that is, applying a different treatment in a designed experiment has an effect on the mean response, or the different skin conditions show differences in their respective mean blood trace amounts. These concepts are defined more precisely in subsequent sections.

In this book, the term **study** refers to whatever type of project is relevant: designed experiment, sample survey, or observational study.

1.2 Statistical Models

Statistical models for data are mathematical descriptions of how the data conceivably can be produced. Models consist of at least two parts: (1) a formula relating the response to all explanatory variables (e.g., effects), and (2) a description of the probability distribution assumed to characterize random variation affecting the observed response.

Consider the experiment with five drugs (say, A, B, C, D, and E) applied to subjects to control blood pressure. Let μ_A denote the mean blood pressure for subjects treated with drug A, and define μ_B, μ_C, μ_D, and μ_E similarly for the other drugs. The simplest model to describe how observations from this experiment were produced for drug A is $Y_A = \mu_A + e$. That is, a blood pressure observation (Y_A) on a given subject treated with drug A is equal to the mean of drug A plus random variation resulting from whatever is particular to a given subject other than drug A. The random variation, denoted by the term e, is called the **error** in Y. It follows that e is a random variable with a mean of zero and a variance of σ^2. This is the simplest version of a **linear statistical model**—that is, a model where the observation is the sum of terms on the right-hand side of the model that arise from treatment or other explanatory factors plus random error.

The model $Y_A = \mu_A + e$ is called a **means model** because the only term on the right-hand side of the model other than random variation is a treatment mean. Note that the mean is also the expected value of Y_A . The mean can be further modeled in various ways. The first approach leads to an effects model. You can define the effect of drug A as α_A such that $\mu_A = \mu + \alpha_A$, where μ is defined as the intercept. This leads to the one-way **analysis of variance** (ANOVA) model $Y_A = \mu + a_A + e$, the simplest form of an **effects model**. Note that the effects model has more parameters (in this case 6, μ and the α_i) than factor levels (in this case 5). Such models are said to be **over-parameterized** because there are more parameters to estimate than there are unique items of information. Such models require some constraint on the solution to estimate

the parameters. Often, in this kind of model, the constraint involves defining μ as the overall mean implying $\alpha_A = \mu_A - \mu$ and thus

$$\sum_{i=A}^{E} \alpha_i = 0$$

This is called a sum-to-zero constraint. Its advantage is that if the number of observations per treatment is equal, it is easy to interpret. However, for complex designs with unequal observations per treatment, the sum-to-zero constraint becomes intractable, whereas alternative constraints are more generally applicable. SAS procedures use the constraint that the last factor level, in this case α_E, is set to zero. In general, for effects models, the estimate of the mean $\mu_A = \mu + \alpha_A$ is unique and interpretable, but the individual components μ and the α_i may not be.

Another approach to modeling μ_A, which would be appropriate if levels A through E represented doses, or amounts, of a drug given to patients, is to use linear regression. Specifically, let X_A be the drug dose corresponding to treatment A, X_B be the drug dose corresponding to treatment B, and so forth. Then the regression model, $\mu_A = \beta_0 + \beta_1 X_A$, could be used to describe a linear increase (or decrease) in the mean blood pressure as a function of changing dose. This gives rise to the statistical **linear regression** model $Y_A = \beta_0 + \beta_1 X_A + e$.

Now suppose that each drug (or drug dose) is applied to several subjects, say, n of them for each drug. Also, assume that the subjects are assigned to each drug completely at random. Then the experiment is a **completely randomized design**. The blood pressures are determined for each subject. Then Y_{A1} stands for the blood pressure observed on the first subject treated with drug A. In general, Y_{ij} stands for the observation on the j^{th} subject treated with drug i. Then you can write the model equation $Y_{ij} = \mu + e_{ij}$, where e_{ij} is a random variable with mean zero and variance σ^2. This means that the blood pressures for different subjects receiving the same treatment are not all the same. The error, e_{ij}, represents this variation. Notice that this model uses the simplifying assumption that the variance of e_{ij} is the same, σ^2, for each drug. This assumption may or may not be valid in a given situation; more complex models allow for unequal variances among observations within different treatments. Also, note that the model can be elaborated by additional description of μ_i—e.g., as an effects model $\mu_i = \mu + \alpha_i$ or as a regression model $\mu_i = \beta_0 + \beta_1 X_i$. Later in this section, more complicated versions of modeling μ_i are considered.

An alternative way of representing the models above describes them through an assumed probability distribution. For example, the usual linear statistical model for data arising from completely randomized designs assumes that the errors have a normal distribution. Thus, you can write the model $Y_{ij} = \mu_i + e_{ij}$ equivalently as $Y_{ij} \sim N(\mu_i, \sigma^2)$ if the e_{ij} are assumed $iid\ N(0, \sigma^2)$. Similarly, the one-way ANOVA model can be written as $Y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$ and the linear regression model as $Y_{ij} \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$. This is important because it allows you to move easily to models other than linear statistical models, which are becoming increasingly important in a variety of studies.

One important extension beyond linear statistical models involves cases in which the response variable does not have a normal distribution. For example, suppose in the drug experiment that c_i clinics are assigned at random to each drug, n_{ij} subjects are observed at the j^{th} clinic assigned to drug i, and each subject is classified according to whether a medical event such as a stroke or heart attack has occurred or not. The resulting response variable Y_{ij} can be defined as the number of subjects having the event of interest at the ij^{th} clinic, and $Y_{ij} \sim \text{Binomial}(\pi_i, n_{ij})$, where π_i is the probability of a subject showing improvement when treated with drug i. While it

is possible to fit a linear model such as $p_{ij} = \mu_i + e_{ij}$, where $p_{ij} = y_{ij}/n_{ij}$ is the sample proportion and $\mu_i = \pi_i$, a better model might be $\pi_i = 1/\left(1 + e^{-\mu_i}\right)$ and $\mu_i = \mu + \alpha_i$ or $\mu_i = \beta_0 + \beta_1 X_i$ depending on whether the effects-model or regression framework discussed above is more appropriate. In other contexts, modeling $\pi_i = \Phi(\mu_i)$, where $\mu_i = \mu + \alpha_i$ or $\mu_i = \beta_0 + \beta_1 X_i$, may be preferable, e.g., because interpretation is better connected to subject matter under investigation. The former are simple versions of logistic ANOVA and logistic regression models, and the latter are simple versions of probit ANOVA and regression. Both are important examples of **generalized linear models**.

Generalized linear models use a general function of a linear model to describe the expected value of the observations. The linear model is suggested by the design and the nature of the explanatory variables, similar to the rationale for ANOVA or regression models. The general function (which can be linear or nonlinear) is suggested by the probability distribution of the response variable. Note that the general function can be the linear model itself and the distribution can be normal; thus, "standard" ANOVA and regression models are in fact special cases of generalized linear models. Chapter 14 discusses mixed model forms of generalized linear models.

In addition to generalized linear models, another important extension involves nonlinear statistical models. These occur when the relationship between the expected value of the random variable and the treatment, explanatory, or predictor variables is nonlinear. Generalized linear models are a special case, but they require a linear model embedded within a nonlinear function of the mean. **Nonlinear models** may use any function, and may occur when the response variable has a normal distribution. For example, increasing amounts of fertilizer nitrogen (N) are applied to a crop. The observed yield can be modeled using a normal distribution—that is, $Y_{ij} \sim N(\mu_i, \sigma^2)$. The expected value of Y_{ij} in turn is modeled by $\mu_i = \alpha_i \exp\{-\exp(\beta_i - \gamma_i X_i)\}$, where X_i is the i^{th} level or amount of fertilizer N, α_i is the asymptote for the i^{th} level of N, γ_i is the slope, and β_i / γ_i is the inflection point. This is a Gompertz function that models a nonlinear increase in yield as a function of N: the response is small to low N, then increases rapidly at higher N, then reaches a point of diminishing returns and finally an asymptote at even higher N. Chapter 15 discusses mixed model forms of nonlinear models.

1.3 Fixed and Random Effects

The previous section considered models of the mean involving only an assumed distribution of the response variable and a function of the mean involving only factor effects that are treated as known constants. These are called **fixed effects**. An effect is called fixed if the levels in the study represent all possible levels of the factor, or at least all levels about which inference is to be made. Note that this includes regression models where the observed values of the explanatory variable cover the entire region of interest. In the blood pressure drug experiment, the effects of the drugs are fixed if the five specific drugs are the only candidates for use and if conclusions about the experiment are restricted to those five drugs. You can examine the differences among the drugs to see which are essentially equivalent and which are better or worse than others. In terms of the model $Y_{ij} = \mu + \alpha_i + e_{ij}$, the effects α_A through α_E represent the effects of a particular drug relative to the intercept μ. The parameters α_A, α_B, ..., α_E represent fixed, unknown quantities.

Data from the study provide estimates about the five drug means and differences among them. For example, the sample mean from drug A, $\bar{y}_{A\cdot}$ is an estimate of the population mean μ_A.

Notation note: When data values are summed over a subscript, that subscript is replaced by a period. For example, $y_{A\cdot}$ stands for $y_{A1} + y_{A2} + ... + y_{An}$. A bar over the summed value denotes the sample average. For example, $\bar{y}_{A\cdot} = n^{-1} y_{A\cdot}$.

The difference between two sample means, such as $\bar{y}_{A\cdot} - \bar{y}_{B\cdot}$, is an estimate of the difference between two population means $\mu_A - \mu_B$. The variance of the estimate $\bar{y}_{A\cdot}$ is $n^{-1}\sigma^2$ and the variance of the estimate $\bar{y}_{A\cdot} - \bar{y}_{B\cdot}$ is $2\sigma^2/n$. In reality, σ^2 is unknown and must be estimated. Denote the sample variance for drug A by s_A^2, the sample variance for drug B by s_B^2, and similarly for drugs C, D, and E. Each of these sample variances is an estimate of σ^2 with n–1 degrees of freedom. Therefore, the average of the sample variances, $s^2 = (s_A^2 + s_B^2 + ... + s_E^2)/5$, is also an estimate of σ^2 with $5(n–1)$ degrees of freedom. You can use this estimate to calculate standard errors of the drug sample means, which can in turn be used to make inferences about the drug population means. For example, the standard error of the estimate $\bar{y}_{A\cdot} - \bar{y}_{B\cdot}$ is $\sqrt{2s^2/n}$. The confidence interval is $(\bar{y}_{A\cdot} - \bar{y}_{B\cdot}) \pm t_\alpha \sqrt{2s^2/n}$, where t_α is the α-level, two-sided critical value of the t-distribution with $5(n–1)$ degrees of freedom.

Factor effects are **random** if they are used in the study to represent only a sample (ideally, a *random sample*) of a larger set of potential levels. The factor effects corresponding to the larger set of levels constitute a population with a probability distribution. The last statement bears repeating because it goes to the heart of a great deal of confusion about the difference between fixed and random effects: *a factor is considered* random *if its levels plausibly represent a larger population with a probability distribution*. In the blood pressure drug experiment, the drugs would be considered random if there are actually a large number of such drugs and only five were sampled to represent the population for the study. Note that this is different from a regression or response surface design, where doses or amounts are selected deliberately to optimize estimation of fixed regression parameters of the experimental region. Random effects represent true sampling and are assumed to have probability distributions.

Deciding whether a factor is random or fixed is not always easy and can be controversial. Blocking factors and locations illustrate this point. In agricultural experiments blocking often reflects variation in a field, such as on a slope with one block in a strip at the top of the slope, one block on a strip below it, and so forth, to the bottom of the slope. One might argue that there is nothing random about these blocks. However, an additional feature of random effects is **exchangeability**. Are the blocks used in this experiment the only blocks that could have been used, or could any set of blocks from the target population be substituted? Treatment levels are not exchangeable: you cannot estimate the effects of drugs *A* through *E* unless you observe drugs *A* though *E*. But you could observe them on any valid subset of the target population. Similar arguments can be made with respect to locations. Chapter 2 considers the issue of random versus fixed blocks in greater detail. Chapter 6 considers the multi-location problem.

When the effect is random, we typically assume that the distribution of the random effect has mean zero and variance σ_a^2, where the subscript a refers to the variance of the treatment effects; if the drugs were random, it would denote the variance among drug effects in the population of drugs. The linear statistical model can be written $Y_{ij} = \mu + a_i + e_{ij}$, where μ represents the mean of all drugs in the population, not just those observed in the study. Note that the drug effect is denoted a_i rather than α_i as in the previous model. A frequently used convention, which this book follows, is to denote fixed effects with Greek letters and random effects with Latin letters. Because the drugs in this study are a sample, the effects a_i are random variables with mean 0 and variance σ_a^2. The variance of Y_{ij} is $\text{Var}[Y_{ij}] = \text{Var}[\mu + a_i + e_{ij}] = \sigma_a^2 + \sigma^2$.

1.4 Mixed Models

Fixed and random effects were described in the preceding section. A **mixed model** contains both fixed and random effects. Consider the blood pressure drug experiment from the previous sections, but suppose that we are given new information about how the experiment was conducted. The n subjects assigned to each drug treatment were actually identified for the study in carefully matched groups of five. They were matched for criteria such that they would be expected to have similar blood pressure history and response. Within each group of five, drugs were assigned so that each of the drugs A, B, C, D, and E was assigned to exactly one subject. Further assume that the n groups of five matched subjects each was drawn from a larger population of subjects who potentially could have been selected for the experiment. The design is a randomized blocks with fixed treatment effects and random block effects.

The model is $Y_{ij} = \mu + \alpha_i + b_j + e_{ij}$, where μ, α_A, ..., α_E represent unknown fixed parameters—intercept and the five drug treatment effects, respectively—and the b_j and e_{ij} are random variables representing blocks (matched groups of five) and error, respectively. Assume that the random variables b_j and e_{ij} have mean zero and variances σ_b^2 and σ^2, respectively. The variance of Y_{ij} of the randomly chosen matched set j assigned to drug treatment i is $\mathrm{Var}[Y_{ij}] = \sigma_a^2 + \sigma^2$. The difference between two drug treatment means (say, drugs A and B) within the same matched group is $Y_{Aj} - Y_{Bj}$. It is noteworthy that the difference expressed in terms of the model equation is $Y_{Aj} - Y_{Bj} = \alpha_A - \alpha_B + e_{Aj} - e_{Bj}$, which contains no matched group effect. The term b_j drops out of the equation. Thus, the variance of this difference is $2\sigma^2/n$. The difference between drug treatments can be estimated free from matched group effects. On the other hand, the mean of a single drug treatment, $\bar{y}_{A\cdot}$ has variance $(\sigma_b^2 + \sigma^2)/n$, which *does* involve the variance among matched groups.

The randomized block design is just the beginning with mixed models. Numerous other experimental and survey designs and observational study protocols produce data for which mixed models are appropriate. Some examples are nested (or hierarchical) designs, split-plot designs, clustered designs, and repeated measures designs. Each of these designs has its own model structure depending on how treatments or explanatory factors are associated with experimental or observational units and how the data are recorded. In nested and split-plot designs there are typically two or more sizes of experimental units. Variances and differences between means must be correctly assessed in order to make valid inferences.

Modeling the variance structure is arguably the most powerful and important single feature of mixed models, and what sets it apart from conventional linear models. This extends beyond variance structure to include correlation among observations. In repeated measures designs, discussed in Chapter 5, measurements taken on the same unit close together in time are often more highly correlated than measurements taken further apart in time. The same principle occurs in two dimensions with spatial data (Chapter 11). Care must be taken to build an appropriate covariance structure into the model. Otherwise, tests of hypotheses, confidence intervals, and possibly even the estimates of treatment means themselves may not be valid. The next section surveys typical mixed model issues that are addressed in this book.

1.5 Typical Studies and the Modeling Issues They Raise

Mixed model issues are best illustrated by way of examples of studies in which they arise. This section previews six examples of studies that call for increasingly complex models.

1.5.1 Random Effects Model

In the first example, 20 packages of ground beef are sampled from a larger population. Three samples are taken at random from within each package. From each sample, two microbial counts are taken. Suppose you can reasonably assume that the log microbial counts follow a normal distribution. Then you can describe the data with the following linear statistical model:

$$Y_{ijk} = \mu + p_i + s(p)_{ij} + e_{ijk}$$

where Y_{ijk} denotes the k^{th} log microbial count for the j^{th} sample of the i^{th} package. Because packages represent a larger population with a plausible probability distribution, you can reasonably assume that package effects, p_i, are random. Similarly, sample within package effects, $s(p)_{ij}$, and count, or error, effects, e_{ijk}, are assumed random. Thus, the p_i, $s(p)_{ij}$, and e_{ijk} effects are all random variables with mean zero and variances σ_p^2, σ_s^2, and σ^2, respectively. This is an example of a **random effects model**. Note that only the overall mean is a fixed effects parameter; all other model effects are random.

The modeling issues are as follows:

1. How should you estimate the variance components σ_p^2, σ_s^2, and σ^2?
2. How should you estimate the standard error of the estimated overall mean, $\hat{\mu}$?
3. How should you estimate random model effects p_i, or $s(p)_{ij}$ if these are needed?

Mixed model methods primarily use three approaches to variance component estimation: (1) procedures based on expected mean squares from the analysis of variance (ANOVA); (2) maximum likelihood (ML); and (3) restricted maximum likelihood (REML), also known as residual maximum likelihood. Of these, ML is usually discouraged, because the variance component estimates are biased downward, and hence so are the standard errors computed from them. This results in excessively narrow confidence intervals whose coverage rates are below the nominal $1-\alpha$ level, and upwardly biased test statistics whose Type I error rates tend to be well above the nominal α level. The REML procedure is the most versatile, but there are situations for which ANOVA procedures are preferable. PROC MIXED in SAS uses the REML approach by default, but provides optional use of ANOVA and other methods when needed. Chapter 4 presents examples where you would want to use ANOVA rather than REML estimation.

The estimate of the overall mean in the random effects model for packages, samples, and counts is $\hat{\mu} = \bar{y}_{...} = \sum y_{ijk} / IJK$, where I denotes the number of packages (20), J is the number of samples per package (3), and K is the number of counts per sample (2). Substituting the model equations yields $\sum (\mu + p_i + s(p)_{ij} + e_{ijk}) / IJK$, and taking the variance yields

$$\mathrm{Var}[\hat{\mu}] = \mathrm{Var}\left[\sum \left(p_i + s(p)_{ij} + e_{ijk}\right)\right] / (IJK)^2 = \left(JK\sigma_p^2 + K\sigma_s^2 + \sigma^2\right) / IJK$$

If you write out the ANOVA table for this model, you can show that you can estimate Var[$\hat{\mu}$] by MS(package)/(IJK). Using this, you can compute the standard error of $\hat{\mu}$ by $\sqrt{\text{MS(package)}/(IJK)}$, and hence the confidence interval for μ becomes

$$\bar{y}_{...} \pm t_{\alpha, df(package)} \sqrt{\text{MS(package)}/(IJK)}$$

where α is the two-sided critical value from the t distribution and df(*package*) are the degrees of freedom associated with the package source of variation in the ANOVA table.

If we regard package effects as fixed, you would estimate its effect as $\hat{p}_i = \bar{y}_{i..} - \bar{y}_{...}$. However, because the package effects are random variables, the **best linear unbiased predictor (BLUP)**

$$\mathrm{E}\left[p_i \mid y\right] = E\left[p_i\right] + \mathrm{Cov}\left[\hat{p}_i, \bar{y}_{i..}\right]\left(\mathrm{Var}\left[\bar{y}_{i..}\right]\right)^{-1}\left(\bar{y}_{i..} - \bar{y}_{...}\right)$$

is more efficient. This leads to the "BLUP"

$$\hat{p}_i = \left(\frac{\sigma_p^2}{\left(JK\sigma_p^2 + K\sigma_s^2 + \sigma^2\right)/JK}\right)\left(\bar{y}_{i..} - \bar{y}_{...}\right)$$

When estimates of the variance components are used, the above is not a true BLUP, but an estimated BLUP, often called an **EBLUP**. Best linear unbiased predictors are used extensively in mixed models and are discussed in detail in Chapters 6 and 8.

1.5.2 Multi-location Example

The second example appeared in Output 3.7 of *SAS System for Linear Models, Fourth Edition* (Littell et al. 2002). The example is a designed experiment with three treatments observed at each of eight locations. At the various locations, each treatment is assigned to between three and 12 randomized complete blocks. A possible linear statistical model is

$$Y_{ijk} = \mu + L_i + b(L)_{ij} + \tau_k + (\tau L)_{ik} + e_{ijk}$$

where L_i is the i^{th} location effect, $b(L)_{ij}$ is the ij^{th} block within location effect, τ_k is the k^{th} treatment effect, and $(\tau L)_{ik}$ is the ik^{th} location by treatment interaction effect. The modeling issues are as follows:

1. Should location be a random or fixed effect?
2. Depending on issue 1, the F-test for treatment depends on MS(error) if location effects are fixed or MS(location × treatment) if location effects are random.
3. Also depending on issue 1, the standard error of treatment means and differences are affected.

The primary issue is one of **inference space**—that is, the population to which the inference applies. If location effects are fixed, then inference applies *only to those locations* actually involved in the study. If location effects are random, then inference applies to the *population represented by the observed locations*. Another way to look at this is to consider issues 2 and 3. The expected mean square for error is σ^2, whereas the expected mean square for location × treatment is $\sigma^2 + k\sigma_{TL}^2$, where σ_{TL}^2 is the variance of the location × treatment effects and k is a

constant determined by a somewhat complicated function of the number of blocks at each location. The variance of a treatment mean is σ^2 / (number of observations per treatment) if location effects are fixed, but it is $[\sigma^2 + K(\sigma_{TL}^2 + \sigma_L^2)]$ / (obs/trt) if location effects are random. The inference space question, then, depends on what sources you believe contribute to uncertainty. If you believe all uncertainty comes from variation among blocks and experimental units within locations, you believe locations are fixed. If, on the other hand, you believe that variation among locations contributes additional uncertainty, then you believe locations are random. Issues of this sort first appear in Chapter 2, and reappear in various forms throughout the rest of the book (e.g., Chapters 4 and 6).

1.5.3 Repeated Measures and Split-Plot Experiments

Because repeated measures and split-plot experiments share some characteristics, they have some modeling issues in common. Suppose that three drug treatments are randomly assigned to subjects, n_i to the i^{th} treatment. Each subject is observed at 1, 2, ..., 7, and 8 hours post-treatment. A possible model for this study is

$$Y_{ijk} = \mu + \alpha_i + s(\alpha)_{ij} + \tau_k + (\alpha\tau)_{ik} + e_{ijk}$$

where α represents treatment effects, τ represents time (or hour) effects, and $s(\alpha)$ represent the random subject within treatment effects. The main modeling issues here are as follows:

1. The experimental unit for the treatment effect (subject) and for time and time × treatment effects (subject × time) are different sizes, and hence these effects require different error terms for statistical inference. *This is a feature common to split-plot and repeated measures experiments.*
2. The errors, e_{ijk}, are correlated within each subject. How best to model correlation and estimate the relevant variance and covariance parameters? This is usually a question specific to repeated measures experiments.
3. How are the degrees of freedom for confidence intervals and hypothesis tests affected?
4. How are standard errors affected when estimated variance and covariance components are used?

Chapter 4 discusses the various forms of split-plot experiments and appropriate analysis using PROC MIXED. Repeated measures use similar strategies for comparing means. Chapter 5 builds on Chapter 4 by adding material specific to repeated measures data. Chapter 5 discusses procedures for identifying and estimating appropriate covariance matrices. Degree of freedom issues are first discussed in Chapter 2 and appear throughout the book. Repeated measures, and correlated error models in general, present special problems to obtain unbiased standard errors and test statistics. These issues are discussed in detail in Chapter 5. Spatial models are also correlated error models and require similar procedures (Chapter 11).

1.5.4 Fixed Treatment, Random Block, Non-normal (Binomial) Data Example

The fourth example is a clinical trial with two treatments conducted at eight locations. At each location, subjects are assigned at random to treatments; n_{ij} subjects are assigned to treatment i at location j. Subjects are observed to have either favorable or unfavorable reactions to the treatments. For the ij^{th} treatment-location combination, Y_{ij} subjects have favorable reactions, or, in other words, $p_{ij} = Y_{ij}/n_{ij}$ is the proportion of favorable reactions to treatment i at location j.

This study raises the following modeling issues:

1. Clinic effects may be random or fixed, raising inference space questions similar to those just discussed.
2. The response variable is binomial, not normal.
3. Because of issue 2, the response may not be linear in the parameters, and the errors may not be additive, casting doubt on the appropriateness of a linear statistical model.
4. Also as a consequence of issue 2, the errors are a function of the mean, and are therefore not homogeneous.

A possible model for this study is a generalized linear mixed model. Denote the probability of favorable reaction to treatment i at location j by π_{ij}. Then $Y_{ij} \sim \text{Binomial}(n_{ij}, \pi_{ij})$. The generalized linear model is

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \mu + c_i + \tau_j + (c\tau)_{ij}$$

or alternatively

$$\pi_{ij} = \frac{e^{\mu+c_i+\tau_j+(c\tau)_{ij}}}{1+e^{\mu+c_i+\tau_j+(c\tau)_{ij}}} = \frac{1}{1+e^{-(\mu+c_i+\tau_j+(c\tau)_{ij})}}$$

where c_i are random clinic effects, τ_j are fixed treatment effects, and $(c\tau)_{ij}$ are random clinic × treatment interaction effects. Generalized linear mixed models are discussed in Chapter 14.

1.5.5 Repeated Measures with Non-normal (Count) Data

The fifth example appears in Output 10.39 of *SAS System for Linear Models, Fourth Edition* (Littell et al. 2002). Two treatments are assigned at random to subjects. Each subject is then observed at four times. In addition, there is a baseline measurement and the subject's age. At each time of measurement, the number of epileptic seizures is counted. The modeling issues here are as follows:

1. Counts are not normally distributed.
2. Repeated measures raise correlated error issues similar to those discussed previously.
3. The model involves both factor effects (treatments) and covariates (regression) in the same model, i.e., analysis of covariance.

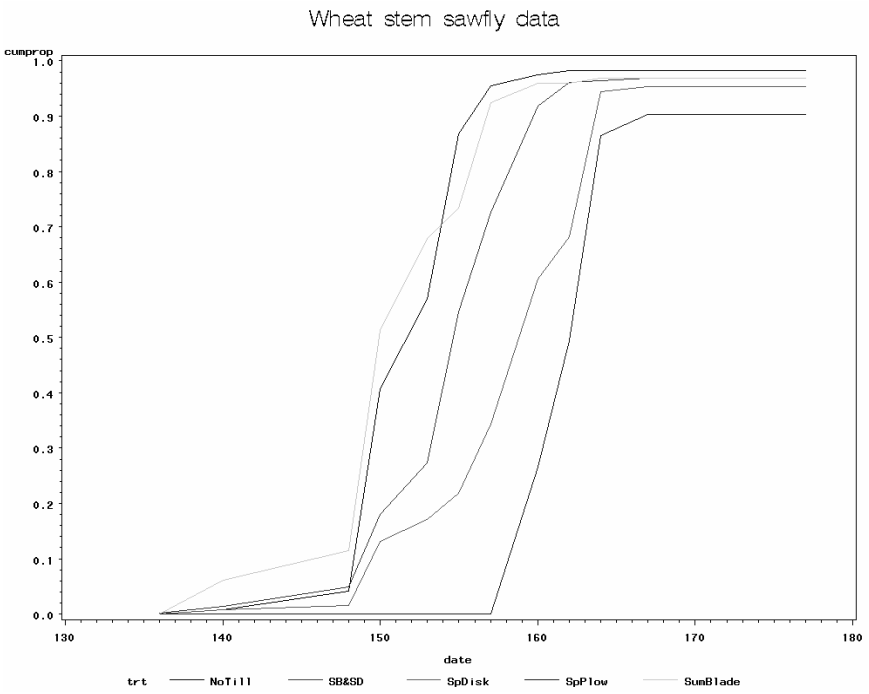
Chapter 7 introduces analysis of covariance in mixed models. Count data in conjunction with repeated measures lead to generalized linear mixed models discussed in Chapter 14.

1.5.6 Repeated Measures and Split Plots with Effects Modeled by Nonlinear Regression Model

The final example involves five treatments observed in a randomized block experiment. Each experimental unit is observed at several times over the growing season and percent emergence is recorded. Figure 1.1 shows a plot of the percent emergence by treatment over the growing season. Like Example 1.5.3, this is a repeated measures experiment, but the structure and model

equation are similar to split-plot experiments, so similar principles apply to mixed model analysis of these data.

Figure 1.1 Treatment Means of Sawfly Data over Time



The modeling issues are as follows:

1. The "usual" mixed model and repeated measures issues discussed in previous examples; plus
2. The obvious nonlinear function required to describe percent emergence as a function of date.

A possible model for this experiment is

$$Y_{ijk} = \mu_{ij} + w_{ij} + e_{ijk}$$

where μ_{ij} is the ij^{th} treatment × date mean, w_{ij} is the random whole-plot error effect, and e_{ijk} are the repeated measures errors, possibly correlated. The Gompertz model described earlier is a suitable candidate to model μ_{ij} as a function of date j for treatment i. The model described here is an example of a nonlinear mixed model. These are discussed in Chapter 15.

1.6 A Typology for Mixed Models

From the examples in the previous section, you can see that contemporary mixed models cover a very wide range of possibilities. In fact, models that many tend to think of as distinct are, in reality, variations on a unified theme. Indeed, the model that only a generation ago was universally referred to as the "general linear model"—fixed effects only, normal and independent errors, homogeneous variance—is now understood to be one of the more restrictive special cases among commonly used statistical models. This section provides a framework to

view the unifying themes, as well as the distinctive features, of the various modeling options under the general heading of "mixed models" that can be implemented with SAS.

As seen in the previous example, the two main features of a statistical model are (1) a **characterization of the mean**, or expected value of the observations, as a function of model parameters and constants that describe the study design, and (2) a **characterization of the probability distribution** of the observations. The simplest example is a one-factor means model where the expected value of the observations on treatment i is μ_i and the distribution is $N(\mu_i, \sigma^2)$, which leads to the linear statistical model $Y_{ij} = \mu_i + e_{ij}$. The generalized linear mixed model from the fifth example of Section 1.5 provides a more complex example: the mean model is

$$\pi_{ij} = 1 \Big/ \left(1 + e^{-(\mu + c_i + \tau_j + (c\tau)_{ij})}\right)$$

and the distribution has two parts—that of the random effects c_j and $(c\tau)_{ij}$, and that of the observations given the random effects, i.e., $Y_{ij} \mid c_j, (c\tau)_{ij} \sim$ Binomial(n_{ij}, π_{ij}). But each model follows from the same general framework.

Appendix 1 provides a more detailed presentation of mixed model theory. In what follows we present an admittedly simplistic overview that uses matrix notation which is developed more fully at appropriate points throughout the book and in the appendix.

Models have two sets of random variables whose distributions we need to characterize: **Y**, the vector of observations, and **u**, the vector of random model effects. The models considered in this book assume that the random model effects follow a normal distribution, so that in general we assume $\mathbf{u} \sim$ MVN(**0**,**G**)—that is, **u** has a multivariate normal distribution with mean zero variance-covariance matrix **G**. In a simple variance components model, such as the randomized block model given in Section 1.4, $\mathbf{G} = \sigma_b^2\mathbf{I}$.

By "mean" of the observations we can refer to one of two concepts: either the **unconditional mean**, E[**Y**] or the **conditional mean** of the observations given the random model effects, E[**Y**|**u**]. In a fixed effects model, the distinction does not matter, but for mixed models it clearly does. Mixed models are mathematical descriptions of the **conditional mean** in terms of fixed effect parameters, random model effects, and various constants that describe the study design. The general notation is as follows:

$\boldsymbol{\beta}$ is the vector of fixed effect parameters.

X is the matrix of constants that describe the structure of the study with respect to the fixed effects. This includes the treatment design, regression explanatory or predictor variables, etc.

Z is the matrix of constants that describe the study's structure with regard to random effects. This includes the blocking design, explanatory variables in random coefficient designs (see Chapter 8), etc.

The mixed model introduced in Section 1.4, where observations are normally distributed, models the conditional mean as $\mathrm{E}[\mathbf{Y}|\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$, and assumes that the conditional distribution of the observations given the random effects is $\mathbf{Y}|\mathbf{u} \sim \mathrm{MVN}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R})$, where **R** is the

variance-covariance matrix of the errors. In simple linear models where errors are independent with homogeneous variances, $\mathbf{R} = \sigma^2\mathbf{I}$. However, in heterogeneous error models (presented in Chapter 9) and correlated error models such as repeated measures or spatial models, the structure of **R** becomes very important.

In the most general mixed model included in SAS, the **nonlinear mixed model** (NLMM), the conditional mean is modeled as a function of **X**, **Z**, β, and **u** with no restrictions; i.e., $h(\mathbf{X}, \mathbf{Z}, \beta, \mathbf{u})$ models $E[\mathbf{Y}|\mathbf{u}]$. Each successive model is more restrictive. The class of **generalized linear mixed models** (GLMM) has a linear model embedded within a nonlinear function—that is, $g(E[\mathbf{Y}|\mathbf{u}])$ is modeled by $\mathbf{X}\beta + \mathbf{Z}\mathbf{u}$. In NLMMs and GLMMs, the observations are not necessarily assumed to be normally distributed. The **linear mixed model** (LMM) does assume normally distributed observations and models the conditional mean directly—that is, you assume $E[\mathbf{Y}|\mathbf{u}] = \mathbf{X}\beta + \mathbf{Z}\mathbf{u}$. Each mixed model has a fixed effects model analog, which means that there are no random model effects and hence **Z** and **u** no longer appear in the model, and the model now applies to $E[\mathbf{Y}]$. The term "mixed model" is often associated with the LMM—it is the "standard" mixed model that is implemented in PROC MIXED. However, the LMM is a special case. The next section presents a flowchart to associate the various models with appropriate SAS software.

Table 1.1 shows the various models and their features in terms of the model equation used for the conditional mean and the assumed distribution of the observations.

Table 1.1 Summary of Models, Characteristics, and Related Book Chapters

Type of Model	Model of Mean	Distribution	Chapter
NLMM	$h(\mathbf{X}, \beta, \mathbf{Z}, \mathbf{u})$	**u**, **Y**|**u** general	15
GLMM	$g^{-1}(\mathbf{X}\beta + \mathbf{Z}\mathbf{u})$	**Y**|**u** general, **u** normal	14
LMM	$\mathbf{X}\beta + \mathbf{Z}\mathbf{u}$	**u**, **Y**|**u** normal	2–11
NLM	$h(\mathbf{X}, \beta)$	**Y** normal	15
GLM	$g^{-1}(\mathbf{X}\beta)$	**Y** general	12
LM	$\mathbf{X}\beta$	**Y** normal	2, 4

1.7 Flowcharts to Select SAS Software to Run Various Mixed Models

SAS offers several procedures (PROCs) designed to implement the various mixed models introduced in the previous sections. PROC MIXED is probably the best known mixed model procedure. It is designed to implement LMMs. SAS has several fixed effects model procedures: PROC GLM implements LMs, PROC NLIN implements NLMs, and PROC GENMOD implements GLMs. There are also several procedures, e.g., LOGISTIC and LIFEREG, that implement special types of GLMs; PROC REG, which implements special types of LMs; and so forth. These special-purpose procedures are not discussed in this book, but they are discussed in detail in other SAS publications as noted throughout this book. Note that PROC GLM was

named before generalized linear models appeared, and was named for "general linear models"; these are now understood not to be general at all, but the most restrictive special case among the models described in Section 1.6, and are now known simply as linear models (LM).

For GLMMs and NLMMs, SAS offers PROC GLIMMIX,[1] PROC NLMIXED, and the %NLINMIX macro. PROC GLIMMIX is the latest addition to the mixed model tools in SAS/STAT. The GLIMMIX procedure fits mixed models with normal random effects where the conditional distribution of the data is a member of the exponential family. Because the normal distribution is also a member of this family, the GLIMMIX procedure can fit LMMs. And because you do not have to specify random effects in the SAS mixed model procedures, PROC MIXED can fit LMs, and PROC GLIMMIX can fit GLMs and LMs. Whereas the GLIMMIX procedure supersedes the %GLIMMIX macro, the %NLINMIX macro continues to have uses distinct and supplementary to the NLMIXED procedure.

Figures 1.2 and 1.3 provide flowcharts to help you select the appropriate model and software for your mixed model project. The basic questions you need to ask are as follows:

- Can you assume a normal distribution for your observations? If the model contains random effects, then this question refers to the conditional distribution of the data, given the random effects.
- Can you assume that the mean or a transformation of the mean is linearly related to the model effects? Note that "linear relation" does not mean the absence of curvature. A quadratic (in X) regression model $\beta_0 + \beta_1 X + \beta_2 X^2$ is a linear model in the β's because all the terms in the model are additive. The linear component is termed the linear predictor. Generalized linear (mixed) models imply such linearity on a certain scale (the transformation $g()$). On the other hand, the Gompertz regression equation (see Sections 1.4 and 1.5) is a nonlinear equation.
- Are all effects (except errors) fixed? Or are there random model effects?
- Can you assume the errors are independent? Or, as in repeated measures or spatial data, are errors possibly correlated?
- A corollary to the previous question is, Are the variances among the errors homogeneous? If the answer is no, then the same modeling strategies for correlated errors are also needed for heterogeneous errors.

Once you answer these questions you can follow the flowchart to see what kind of model you have and what SAS procedure is appropriate. Then you can refer to the relevant chapter in this book for more information about the model and procedures.

[1] The GLIMMIX procedure is an add-on in SAS 9.1 to SAS/STAT for the (32-bit) Windows platform. It does not ship with SAS 9.1. You can obtain the GLIMMIX procedure for SAS 9.1 as a download from www.sas.com/statistics. This site also contains the documentation for the GLIMMIX procedure.

Figure 1.2 Flowchart Indicating Tools in SAS/STAT for Normal Distributed Response

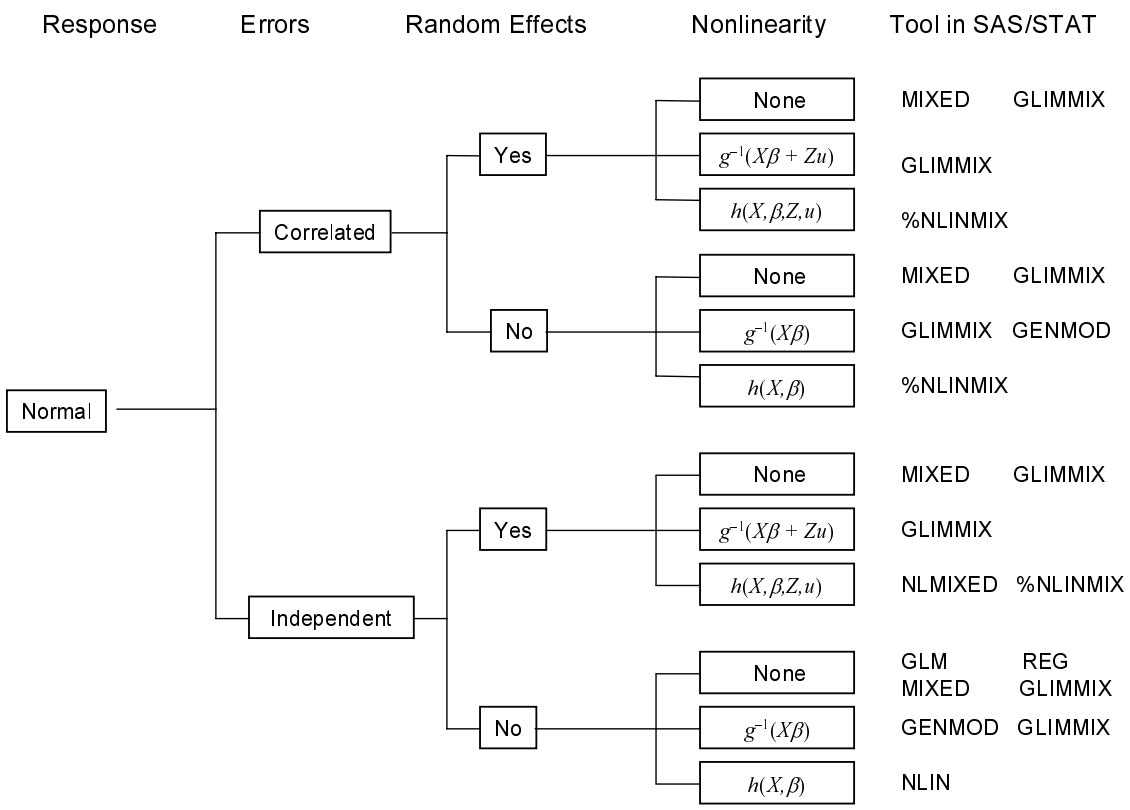


Figure 1.3 Flowchart Indicating Tools in SAS/STAT for Non-normal Distributed Response