Paper 325-2011

The Evolution of Linear Models in SAS: A Personal Perspective

Ramon C. Littell, Info Tech Inc, Gainesville, FL

ABSTRACT

Phenomenal growth in computational power from 1970 through 2010 enabled a parallel expansion in linear model methodology. From humble beginnings in agriculture, linear model applications are now essential in sciences of genetics, education, and biostatistics, to name a few. Indeed, the meaning of "linear models" has evolved accordingly. Developers at SAS Institute have been in the forefront of invention and implementation of these methods at the core of statistical science. Pathways will be traced in steps of SAS® procedures, beginning with GLM and REG, proceeding through VARCOMP, NLIN, MIXED and GENMOD, and arriving at NLMIXED and GLIMMIX. Along the way, some problems have disappeared, new ones have emerged, and others are still along for the ride.

INTRODUCTION

The purpose of this paper is to chronicle the evolution of linear models in SAS® from the perspective of an outsider who has closely followed the progression and whose professional career was influenced by it. Linear models have been in the core of statistical methodology and SAS procedures followed that pattern.

The year 1976 can be considered the birth date of SAS as we now recognize it. SAS·76 was the first release of SAS Incorporated. So one may think of time since 1976 as the Common Era of SAS. The hallmark statistical procedure in SAS·76 was GLM. It was highly innovative for its time and caught attention of statisticians and others engaged in data analysis across the US and beyond. GLM established a pattern for statistical procedures in SAS. Instead of a large number of special purpose linear model applications, GLM provided a comprehensive platform that enabled a user to obtain solutions for most problems falling in the arena of linear models; for regression analysis, analysis of variance and covariance, and multivariate analysis. Whereas most of the capabilities of GLM were inspired by statisticians working in agriculture research, GLM became the workhorse procedure for pharmaceutical statisticians and biostatisticians.

A few years later the REG procedure was released. It expanded regression capabilities to include diagnostic techniques that had been the subject of active research, and recently published in a major text book by Belsley, Kuh and Welsch (1980). Now the user not only had the capability to compute inferential statistics in regression analysis, but could also obtain statistics to help decide what variables to include in the analysis and to identify problematic data.

The VARCOMP procedure provided estimates of variance components in mixed linear models, giving the user four choices of methods of estimation that have also been incorporated into later SAS procedures. This procedure, like GLM, brought forth computing machinery that opened the door to evaluation and comparison of statistical methods which were previously infeasible.

The NLIN procedure, although not really intended for linear models, permitted the formulation of models with linear components, such as segmented polynomials, as nonlinear models.

Capabilities for analysis of categorical data were limited in early versions of SAS. They were enhanced by the CATMOD and GENMOD procedures. CATMOD was based on methodology of Grizzle, Starmer and Koch (1969) that innovated using linear models for categorical data analysis. A later procedure GENMOD was based on generalized linear models introduced by Nelder and Wedderburn (1972).

During the 1980's GLM added useful enhancements, but was nagged by the need for features to adequately accommodate problems related to analysis of correlated data. The immensity of this need inspired the development of the MIXED procedure. Now data with random effects and repeated measures could be analyzed by incorporating those features into the statistical model for the data. Whereas GLM was built around the model for the expected value of the response variable taking all independent variables as fixed, MIXED is built around models for both the expected value of the response as a function only of the fixed variables, and the variance of random effects. This turned the tables in the relation between statistical methodology and its computational implementation. MIXED revealed the need for further development of methods to adjust for the effects of using variance estimates in place of true variances

Shortly following MIXED, macros were provided for fitting nonlinear mixed models and generalized linear mixed models using MIXED to make iterative computations. These macros later evolved into the procedures NLMIXED and GLIMMIX. The GLIMMIX procedure extends the capabilities of GLM and MIXED to generalized linear models.

SAS·76: THE GLM PROCEDURE

The GLM procedure is the first in a line of SAS statistical procedures that have a similar syntax for defining a model and various options. It was released in what was to be known as SAS·76 (SAS Institute, 1976), the first commercial release by SAS Institute.

GLM is essentially a regression procedure utilizing least squares to fit the model. You would fit the following statistical model, with three independent variables x_1, x_2, x_3, and dependent variable y,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e,$$

by using the statements

```
proc glm;
model y = x1 x2 x3;
run;
```

Results printed in the output include parameter estimates and other regression computations, including an analysis of variance and associated statistics. The parameter estimates yield the prediction equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3.$$

Standard errors, t-statistics and p-values are automatically printed. These statistics assume the probability distribution

$$e \sim NID(0, \sigma^2).$$

The analysis of variance partitions the total variation into the portions associated with the independent variables and that not associated with the independent variables,

$$SSTotal = \sum (y - \bar{y})^2 = SSModel + SSError = \sum (y - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2.$$

When it is necessary to more precisely describe which variables are included in the model, we write

$$SSTotal(x_1, x_2, x_3) = SSModel(x_1, x_2, x_3) + SSError(x_1, x_2, x_3).$$

And if it is necessary to clarify that there is an intercept in the model, we write

$$SSTotal(x_1, x_2, x_3 \mid \text{int}) = SSModel(x_1, x_2, x_3 \mid \text{int}) + SSError(x_1, x_2, x_3 \mid \text{int}).$$

Equivalently, it may be convenient to write

$$SSTotal(\beta_1, \beta_2, \beta_3 \mid \beta_0) = SSModel(\beta_1, \beta_2, \beta_3 \mid \beta_0) + SSError(\beta_1, \beta_2, \beta_3 \mid \beta_0).$$

This fundamental idea of partitioning variation is useful for obtain a conceptual, if not mathematical, understanding of some of the many computations brought forth by GLM. In particular, it yields the meaning of variation associated with a set of variables *adjusted for* (or *controlling for*) another set of variables. Specifically, the variation due to x_2, adjusted (controlled for) x_1, is

$$\begin{aligned} SS(\beta_2 \mid \beta_1) &= SSModel(\beta_1, \beta_2 \mid \beta_0) - SSModel(\beta_1 \mid \beta_0) \\ &= SSError(\beta_1 \mid \beta_0) - SSError(\beta_1, \beta_2 \mid \beta_0) \end{aligned}$$

The quantity $SS(\beta_2 \mid \beta_1)$ is also called the *reduction in error sum of squares* due to adding the variable x_2 to the model that already contains the variable x_1 (and only the variable x_1).

The basic feature that gives GLM power beyond ordinary regression is the CLASS statement, which provides automatic computation of indicator variables corresponding to the levels of a classification variable. The indicator variables enable GLM to compute, for example, sums of squares for classification variables and combinations of classification and continuous variables, and perform test of hypotheses. This is the source of the word "general" in the acronym GLM, which was entirely appropriate in 1976 when GLM was released. The landmark paper on general*ized* linear models by Nelder and Wedderburn (1972) was not yet widely known. Since then, however, there has been confusion between "general linear models" and "generalized linear models." And later, there were mixed models, linear mixed models, and generalized linear mixed models. Even yet, universally accepted terminology is lacking. But for now, let's look closely at innovations that appeared in GLM, because they are also found in the later methodologies.

One on the innovations that received a lot of attention after GLM was released was that of the different types of sums of squares, of which there are four. These are labeled Type I, Type II, Type III, and Type IV in GLM. The concepts of two types of sums of squares, sequential and partial, were well know and straightforward to understand from a regression perspective. In terms of the reduction in error sum of squares notation, the sequential and partial sums of squares for the variables are:

Variable	Sequential	Partial
x_1	$SS(\beta_1 \mid \beta_0)$	$SS(\beta_1, \beta_2, \beta_3 \mid \beta_0)$
x_2	$SS(\beta_1, \beta_2 \mid \beta_0)$	$SS(\beta_1, \beta_2, \beta_3 \mid \beta_0)$
x_3	$SS(\beta_1, \beta_2, \beta_3 \mid \beta_0)$	$SS(\beta_1, \beta_2, \beta_3 \mid \beta_0)$

The Type I sum of squares is always sequential, and depends on the order of variables specified in the model statement. In a ordinary regression context, Types II, III and IV are all partial. Distinctions between Types II, III and IV occur when class variables are specified. Then these latter three types may all differ, but time and space do not permit a complete description. See Littell, Stroup and Freund (2005), Hocking, Hackney and Speed (1978), and Searle (1987) for details. In brief, the topic can be illustrated in terms of a two-way cross-classification of data, with factors A and B, that have a and b levels, respectively. Then a model would be

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk},\ i = 1,...,a;\ j = 1,...,b;\ k \le n_{ij},$$

where n_{ij} is the number of observations corresponding to levels i of A and j of B. Types III and IV will be the same if $n_{ij} > 0$ for all i, j. Probably the most popular way to try to interpret the different types of sums of squares for a factor is in terms of the "hypothesis tested" by the F statistic derived based on the sum of squares for the factor. To describe further, denote $E(y_{ijk}) = \mu_{ij}$, which represents the "population cell mean" corresponding to levels i of A and j of B. . Then, in term of the model parameters, μ_{ij} is represented as

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}.$$

A statistic about A would have null hypothesis of the form

$$H_0 : \tilde{\mu}_{1.} = ... = \tilde{\mu}_{a.}.$$

where $\tilde{\mu}_{i.} = \sum_i w_{ij}\mu_{ij}$ is a weighted average of the cell means. The different types of sums of squares for factor A are determined by the values of the weights, w_{ij}. These can be exceeding complicated to obtain and comprehend in many situations. But if $n_{ij} > 0$ for all cells, then Types III and IV will be equal, and the weights are $w_{ij} = 1/b$ for all i, j. If $n_{ij} = 0$ for one or more cells, then Types III and IV typically differ.

The question of practical importance is whether one can obtain a test of the desired hypothesis; or, more fundamentally, whether a meaningful hypothesis can be specified. With so-called unbalanced data (or worse, with empty cells), this is not a simple matter. Arguments and discussions abounded during the late 1970's and the 1980's.

It is important recognize that these concepts apply as well in M IXED, GLIMMIX other procedures. Unfortunately, this

issue seems to be more forgotten with each decade, as statistical methods become more complex and less attention is paid to fundamental concepts. Focus these days is as much on computation as statistical inference.

Prior to about 1970, there were almost no statistical computing packages generally available. Individual institutions developed their own software, with varying degrees of quality. Computational power was in its infancy, and if one was able to obtain any feasible looking answer, it was seldom challenged. Only in the 1970's did statistical computing packages become commonplace, and it became apparent that they did not all provide the same computations for the same questions of the same data. This is the setting in which GLM was designed and written.

Along with questions of what computations should be made (e.g., what sums of squares to compute) were questions of how to approach the computations. This is where GLM was on the forefront. Certain specifications of statistical models were too restrictive to design a procedure that would accommodate a vast array of problems,. In particular, standard assumptions in textbooks such as $\sum_i \alpha_i = 0$ or $\alpha_a = 0$ in order to obtain full-rank models were limiting. Instead of building in such assumptions, GLM presumes no artificial restrictions on the parameters. Parameter estimates are obtained through generalized inverses. But then interpretations of parameters may be obscure, so the so-called estimable functions were made available for almost all linear computations. In principle, one could determine "what hypothesis was being tested" from the estimable functions.

Numerous technical problems have been encountered by SAS developers. How to implement the model in GLM must have been one of the most challenging. James H. Goodnight, who designed and wrote all of the early versions of GLM, chose the road less traveled and cleared the way for many developers to follow. A notable contribution in the statistical literature is his paper on the sweep operator (Goodnight, 1979). Anyone who wonders how the parameter estimates for classification variables are computed will be enlightened by the paper.

In parallel with computations of sums of squares were computations of means for levels of a classification variable. Corresponding to the weighted means in the test of hypothesis are various ways the means could be formulated; i.e., how averages should be computed across level of another factor. GLM provided three statements for obtaining means. One is the MEANS statement which computes mean and other statistics for combinations of classification variables without regard to other variables. A second is the LSMEANS statement, which computes estimates of the expected values defined by the model, and, in fact, are referred to as "model-based" means in certain literature. Basically, weights were assigned that gave equal values to levels of the other factor, subject to estimability. In that sense, the LSMEANS correspond to the Type III sums of squares. In other situations, typically where there are empty cells, estimability breaks down for LSMEANS. Types III and IV sums of squares are available in all situations, but their interpretation may be suspect where there are empty cells. The other option offered by GLM is the ESTIMATE statement (and its companion, the CONTRAST statement) which permits the user to specify any estimable function desired. These devices are extremely powerful and flexible, but are difficult to use without a good knowledge of their inner works.

Again, LSMEANS are computed by both MIXED and GLIMMIX with possibly different meaning from GLM. One may consider the term "LSMEANS" inappropriate for the procedures MIXED and GLMMIX because these procedures technically are not "least squares" the usual sense.

Historically, the term LSMEANS came from the field of animal breeding, and was introduced to the broader statistical community through SAS via Walter Harvey and Bill Sanders. Harvey (1975) wrote a remarkable technical bulletin for its time describing least squares computational techniques that were implemented into GLM, including "absorption," which permits computation of a subset of regression parameters in a model with a large number of variables without completely solving the normal equations. This technique is implemented in the ABSORB statement in GLM due to influence of animal breeders who deal with large numbers of observations and variables. It also is in the package SYSTAT, which is popular among econometricians. I speculate that knowledge of the technique may have come from its implementation in SAS. This is just one of many advances in statistical methodology and computing that came from a community of statisticians who consider themselves primarily as members of another profession.

The concept of LSMEANS is similar to the concept for Type III hypotheses; to average across levels of other factors using equal weights. For example, the LSMEAN for level 1 of factor A is

$$\mu_{1.} = \sum_j \mu_{1j} / b = \mu + \alpha_1 + \sum_j \beta_j / b + \sum (\alpha\beta)_{1j} / b = \mu + \alpha_1 + \overline{\beta} + (\alpha\beta)_{1.}$$

The other way to obtain means (or any other estimable linear combination) is to use the ESTIMATE statement. To illustrate, assume a=3 and b=5. You can duplicate the LSMEAN using the statement

```
ESTIMATE intercept  1  A 1 0 0  B .2 .2 .2 .2 .2  A*B .2 .2 .2 .2 .2  0 0 0 0 0  0 0 0 0 0;
```

The ESTIMATE statement is available in MIXED and GLIMMIX in essentially the same syntax, but with a vast array

of options.

The power and generality of SAS procedures, like everything else, comes at a cost to the user. Perhaps the most prominent illustration of the extracted cost is the necessity of dealing with the concept of estimability, which arises because of utilizing singular systems of normal equations and getting solutions using generalized inverses. The cost to the user is learning enough to be fluent in the technical aspects. In terms of the ESTIMATE statement above, this means knowing how and where to place the coefficients on model parameters. But the benefit to the user is doing it correctly rather than guessing and hoping it's right.

GLM was the first SAS procedure in the Common Era of SAS to explicitly provide applications that accommodate random effects. Suppose that factor B in the model above is random rather than fixed. Then the model might be written

$$y = \mu + \alpha_i + b_j + (\alpha b)_{ij} + e_{ijk},$$

where $b_j \sim NID(0, \sigma_b^2)$ and $(\alpha b)_{ij} \sim NID(0, \sigma_{\alpha b}^2)$.

This is specified using the RANDOM statement

```
RANDOM B A*B;
```

The basic output of the RANDOM statement in GLM is a table of expected means squares, which allow the user to determine, among other things, the appropriate denominators for F statistics. The probability distribution of random effects (shown above) is incorporated into the expected mean square computations of GLM. Users who have studied expected mean squares in the classical textbooks have been confused because the expected mean squares computed by GLM do not agree with those they know from the classical texts. This issue, not unlike to one regarding computations of sums of squares, caused a flurry of attention. Basically, there are alternative ways to prescribe the model. This issue is discussed in Hocking (1987) and Littell, Stroup and Freund (2000), but it is still unresolved how to correctly specify a model for random effects for a given application.

Like the LSMEANS and ESTIMATE statement, the RANDOM statement is available in MIXED and GLIMMIX. Its purpose is essentially the same, but technical differences exist. The dilemma of how to define the random effects in GLM carries forward in MIXED and GLIMMIX, but receives little or no attention.

As noted earlier, estimability in GLM is judged on the basis of all effects in the model being fixed, even if some are listed in the random statement. This often results in judging a linear combination of parameters as being non-estimable when in fact it is estimable in concert with the random effects. For example, if there are empty cells in the example of the A-by-B two-way classification, A is fixed and B is random, then LSMEANS for A would be considered non-estimable by GLM, when they are theoretically estimable. Practical examples where this occurs are in randomized block design, multi-center clinical trials, and cross-sectional studies where blocks, clinics, and sections are considered random, respectively. One of the major advancements from GLM to MIXED and GLIMMIX is that computational machinery in MIXED and GLIMMIX is based on modeling framework that builds in the random effects, which will be discussed more thoroughly in the section on MIXED. First, we take a look at other procedures that were also introduced in SAS·76.

VARIABLE SELECTION PROCEDURES: STEPWISE AND RSQUARE

Two procedures for variable selection in regression are STEPWISE and RSQUARE. STEPWISE provided five options of rules for variable selection based on forward selection, backward elimination, a combination of thereof, and R^2 improvement. RSQUARE gave variable names and the value R^2 for all possible models. These procedures were powerful tools for selecting variables for a regression equation, which could then be fitted with GLM.

ECONOMETRIC PROCEDURES: AUTOREG AND SYSREG

Two procedures with roots in econometrics are AUTOREG and SYSREG. AUTOREG fits autoregressive models and allows the user to specify the order of lags. SYSREG fits interdependent systems of linear equations. These two procedures extended the community of SAS users beyond the realm of experimental statistics.

OTHER PROCEDURES: NLIN AND VARCOMP

The NLIN and VARCOMP procedures were included in SAS·76. Both were substantially enhanced in later versions. NLIN, of course, is not really a linear models procedures, but is useful for fitting models with linear components, such as so-call linear plateau models. Such models consist of a line segment from, say x=a to x=b, and another line segment from x=b to x=c with slope=0 and joining the first segment at x=b, where b in not known. NLIN can be used to fit this model. NLIN also contained a feature not found in other SAS statistical procedures. It essentially had a built in programming language that allowed the user to define functions of variables in mathematical equation form.

VARCOMP provided variance component estimates based on the expected mean squares from and analysis of variance. Later versions of VARCOMP implemented contemporary method of estimation.

ADVANCEMENTS IN LATE 1970'S AND 1980'S

During the period of time from 1976 to 1990 some new statistical procedures were released, notably REG and CATMOD. In addition, GLM and VARCOMP were enhanced to accommodate computations that were previously not available.

REPEATED MEASURES IN GLM

Analysis of repeated measures data was becoming increasingly important, due in part to the expansion of drug evaluations over time in the pharmaceutical industry. Repeated measures is a topic that was developed largely in the fields of psychology and education, stemming from the use of repeated testing in these fields. Multivariate analysis in GLM could be used to perform certain analyses of repeated measures data, taking the sequence of repeated measures on each subject as a multivariate vector of data. In SAS, that meant recording all the repeated measures in a single OBS. The REPEATED statement in GLM essentially automated several of these multivariate computations, and presented output in the context of repeated measures terminology. In addition, the REPEATED statement implemented methods to assess the degree of departure of the covariance of the repeated measures data from the structure necessary for straight-up analysis of variance methods (Huynh and Feldt, 1970), and made adjustments to significance probabilities that were calculated from an analysis of variance. These adjustments are useful, but do not substitute for the methods in MIXED, which gives the user a means of formulating the actual covariance according to an specified structure. The REPEATED statement in GLM established terminology that was carried forth in later procedures; the terms REPEATED and SUBJECT, which have broader meaning than the words imply.

RANDOM EFFECTS IN GLM

The TEST statement was available in SAS·76, which allowed the user to specify both the numerator and denominator mean square in a F test. The RANDOM statement presented a table of expected mean squares that enabled the user to determine an appropriate denominator for the test, if one existed directly. The TEST option on the RANDOM caused GLM to identify and compute linear combinations of mean squares to form an appropriate denominator, and also computed approximate degrees of freedom based on Satterthwaite's formula. This was done for all effects in the MODEL statement. In addition, the user could specify a linear combination of effects to test using a CONTRAST statement. GLM gave an expected mean square for the linear combination (which, by the way, was not limited to contrasts). A creative user could use this facility to compute an appropriate standard error by hand for any linear combination of effects specified in an ESTIMATE statement. This capability was never built into GLM, even though the analogous method for testing was an option in the RANDOM statement.

VARIANCE COMPONENT ESTIMATION

The concept of random effects has been familiar to students of statistics (at least in the land grant universities) for several decades. Models such as the one from the previous section,

$$y = \mu + \alpha_i + b_j + (\alpha b)_{ij} + e_{ijk}$$

are discussed in some of the earliest test books; e.g. Steel and Torrie (1960) and Snedecor and Cochran (5th ed, 1967). But estimation of the variance components (σ_b^2 and $\sigma_{(\alpha b)}^2$) was generally limited to analysis of variance derived from the expected mean squares. By 1976 several other methods had been developed but not implemented in statistical data analysis systems. The VARCOMP procedure, which contained in SAS·76, received major later enhancements in computational capability. It allowed the user to choose from four methods, called Type1, MIVQUE0, ML, and REML. Type1 is the analysis of variance method, based on the Type 1 expected mean squares.

MIVQUE0 is a method related to Type1, but adjusts mean squares only for fixed effects, thereby being computationally more efficient. ML is the maximum likelihood method of estimating the variance components. REML (Patterson and Thompson, 1971) is maximum based on the residuals from fitting a model with only fixed effects. The residuals thereby are functions only of random effects. REML has become the method of choice of most statisticians. Although comprehensive optimality properties have not been established, REML is has known attractive features. For example, REML estimates are less biased than ML, and in many cases are unbiased. The classical example is computation of the variance of a single sample of data. One obtains an unbiased estimate of the population variance by using the denominator n-1, which is REML. The ML estimate has n in the denominator, and is biased by a factor of (n-1)/n. In broader applications, ML estimates of variances lead to inflated test statistics and optimistic confidence intervals. Swallow and Monahan (1984) made a comprehensive comparison of methods for estimating variance components, and therein established a protocol for simulation studies in statistics.

THE REG PROCEDURE

For strictly regression applications, the most important advancement during the 1980's was the REG procedure, designed and written by John Sall. REG had built-in applications for variable selection. It also implemented the vast array of diagnostic techniques described by Belsley, Kuh and Welsch (1980). It immediately jumped to the front of all available regression programs, as it contained essentially every known method for models with homoscedastic and independent errors. In addition, REG has facilities for variable selection. To this day, REG is unsurpassed in ease of use, wealth of features, and computational efficiency.

THE CATMOD, GENMOD, AND LOGISTIC PROCEDURES

The CATMOD procedure is based on the landmark paper by Grizzle, Starmer and Koch (1968), which gave a linear model approach to analysis of categorical data. It assumed that one or more of the categorical variables may be considered response variables, and the others independent variables. It models functions the outcomes of the dependent variable as linear functions of the independent variables, and permitted analysis of variance-type inference about the factors.

The GENMOD procedure is based on the methods proposed by Nelder and Wedderburn (1972). It also assumes one of the variables plays the role of a dependent variable that is to be related to another set of variables. Any of the variables may be continuous or discrete. The modeling approach set the stage for a revolution in methods for analyzing data. It proposes a regression-type model which relates the expected value of the dependent variable to a function of a linear combinations of the other variables. Commonly used notation specifies

$$\eta = g(\mu) = x'\beta \text{ and } E(y) = \mu = h(x'\beta) .$$

The functions g and $h = g^{-1}$ are called the "link" and "inverse link" functions, respectively.

called a "generalized linear model." Certain conditions are assumed about the conditional distribution of $y \mid x$ and properties of the link function that make maximum likelihood estimates feasible to compute.

The term "generalized linear model" is often confused with "general linear model." The former is more general, and was given the acronym GLIM. This held up for several years, but eventually generalized linear models laid claim to the name GLM. If everything were renamed, GLM would be changed to LM and GLIM to GLM. In this paper, the earlier meaning of GLM and GLIM will continue to be used avoid confusion. GLM is actually a special case of GLIM, with the link function being the identity function and the distribution of $y \mid x$ normal.

The LOGISTIC procedure implements logistic regression, which also is a special case of the generalized linear model, with link function given by the inverse of the logistic distribution function and Bernoulli distribution for $y \mid x$.

THE 1990'S

Most statisticians would consider the advent of the MIXED procedure to be the seminal event in SAS/STAT of the last decade of the 20th century. The procedure was released in 1992, but has a history long before that. In previous years, users were trying to make GLM perform mixed model computations that went beyond the basic capabilities of the procedure. Some of these could be obtained by manipulating GLM and exploiting the RANDOM statement, but obtaining some standard errors were beyond practical hope. Demands for high-quality analyses of repeated measures and split plot (hierarchical) data were principle drivers. At the head of the pack of such users were statisticians representing land grand university agricultural experiment stations. One group in particular, known informally as University Statisticians of Southern Experiment Stations (USSES) has been in place since at least the

early 1960's. USSES members collaborated on many projects over the years, and in the late 1980's developed software and authored a publication ***Applications of Mixed Models in Agriculture and Related Disciplines***, Southern Cooperative Series Bulletin No. 343, Louisiana Agricultural Experiment Station (1989). A meeting of USSES was hosted by SAS Institute, and shortly after that meeting SAS embarked on development of MIXED. Documentation of MIXED refers to articles in this bulletin, e.g. Stroup (1989) and Giesbrecht (1989). Articles of other members of USSES are also cited in MIXED documentation, notably (McLean and Sanders (1988) and McLean, Sanders and Stroup (1992), Giesbrecht and Burns (1985), and Fai and Cornelius (1996).

It can fairly be said that one person, Bill Sanders, had more influence on the development of GLM and MIXED than anyone else outside of SAS Institute. Undoubtedly, his efforts have had an immense impact on the use of statistics. This is all the more remarkable because he does not hold a degree in statistics, but rather in animal science. It may not be widely known that some of the most important work in applied linear models came from such people, notably Shayle Searle, David Harville, Walter Harvey, Charles Henderson, Oliver Schabenberger, and many others. Sanders may not be well-known among card-carrying statisticians, but he has surely affected their ability to perform statistical analyses. During the past two decades, Sanders has focused on using mixed models to evaluate student, teacher and school achievement, based on the general concept of best linear unbiased prediction (Henderson, 1984), and has achieved national stature in that arena.

THE MIXED PROCEDURE

The MIXED procedure was the most hailed statistical development at SAS Institute in the 1990's. It brought forth mixed model methodology that was only a dream a decade earlier, and in my opinion, resulted from a perfect storm of need and input from users combined with technical capability, resources and commitment from SAS Institute. MIXED is one more example of an humble idea finding its way to great a product. Russell Wolfinger designed and wrote MIXED. Related to MIXED, he has conducted seminal research and published widely in the statistical literature.

The basic statistical method implemented in MIXED is based on generalized least squares. The statistical model is

$$Y = X\beta + Z\gamma + e, \tag{1}$$

where Y is a vector of data, β is a vector of fixed effect parameters, γ is vector of random effects, and e is a vector of errors. The random vectors are assumed to have the distributions $\gamma \sim N(0, G)$ and $e \sim N(0, R)$, and are independent of each other. Thus Y has the distribution

$$Y \sim N(X\beta, ZGZ' + R).$$

The model could be equivalently specified

$$Y = X\beta + \varepsilon, \tag{2}$$

where $\varepsilon \sim N(0, V)$ and $V = ZGZ' + R$.

MIXED and GLM have similar syntax, with important distinctions. Most importantly, in MIXED you specify the fixed effects in the MODEL statement and the random effects in RANDOM and/or REPEATED statements. Thus, the syntax reveals the essential distinction between the GLM and MIXED procedures. In MIXED, the fixed effects and the random effects are formulated separately, whereas in GLM the fixed effect and random effects are all included in the same model statement.

The generalized least squares (GLS) estimate of β is

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

In many situations, the inverses do not exist, and generalized inverses must be used, but we shall not get into the ramifications of that. The covariance matrix of $\hat{\beta}$ is $V(\hat{\beta}) = (X'V^{-1}X)^{-1}$. But therein lies the rub: V must be known in order to estimate β and compute the covariance matrix of the estimate. This is hardly ever the case in reality. In certain special cases, such as a balanced randomized blocks design, $\hat{\beta}$ can be computed using ordinary least squares, $\hat{\beta} = (X'X)^{-1}X'Y$, but even then, the covariance matrix will require computation of V. This brings us to the (usually) most difficult step in using the MIXED procedure; specifying the form of V. This is what you do with the RANDOM and REPEATED statements. The RANDOM statement defines G and the REPEATED statement defines R. The two statements may be used singly or together.

For example, consider the mixed model you saw previously,

$$y = \mu + \alpha_i + b_j + (\alpha b)_{ij} + e_{ijk},$$

where $b_j \sim NID(0, \sigma_b^2)$ and $(\alpha b)_{ij} \sim NID(0, \sigma_{\alpha b}^2)$.

You would use the statements

```
PROC MIXED; CLASS A B;
MODEL  Y=A;
RANDOM B A*B;
RUN;
```

The RANDOM statement looks like the one you saw for GLM, but has greatly different function.

You can think of this code as corresponding to the model description (1).

Fixed effect parameter estimates, standard errors, tests of hypotheses are all computed by inserting the estimate $\hat{V}$ in place of V into the GLS formulas. Sometimes the estimates are called “estimated” or “empirical” GLS, (EGLS). Generally speaking, the EGLS estimates are unbiased, but their variances are inflated due to the variation in $\hat{V}$.

Relatively little was known about the consequence of the substitution of $\hat{V}$ for V when MIXED first appeared. Later versions of MIXED invoked methods due to Prasad and Rao (1990), Kackar and Harville (1984), and Kenward and Rogers (1997).

In addition to adjustments to the standard errors, degrees of freedom require adjustments. Methods due to Giesbrecht and Burns (1986) and Kenward and Roger (1997) are used in MIXED.

There is good news and bad news regarding the non-estimability problem in GLM when one moves to MIXED. The bad news is that it’s still there. The good news is that the problem is diminished due to estimability being judged only in relation to the fixed effects. Likewise, the difficulty with deciding among the four times of sums of squares in GLM persists, but only in relation to the fixed effects.

New problems in MIXED include effects of different types of variance estimates on EGLS estimates. Is it better to use ML or REML? Should estimates of “non-significant” variances be retained in estimation of fixed effects?
ones have emerged, and others are still along for the ride.

EXAMPLE TO COMPARE MIXED AND GLM IN A MULTI-CENTER CLINICAL TRIAL

Side effects of two drugs were investigated in a multi-center clinical trial. Patients at fifty-three clinics were randomized to the drugs. Following administration of the drugs, patients returned to the clinics at five tri-weekly visits. At each visit, several clinical signs were recorded, including sitting heart rate, (si_hr). The numbers of patients on each drug at each clinic ranged generally from one to ten, although there were no patients on one or the other drug at a small number of clinics. Also, there were more than ten patients on each drug at two clinics. Clinics (which are designated “inv,” abbreviating “investigator”) are considered random because it is desired to make inference applicable to a broader population of clinics. Also, patients are considered random to represent samples of patients from the populations of patients at each clinic. In addition, there is residual variation at each visit for each patient.

Let y_{ijkl} be the measure of sitting heart rate (si_hr) at time *l* on patient *k* on drug *i* at clinic *j*. When developing a statistical model for the data, it is helpful to imagine the sources of sampling variation as if drugs had not been assigned. These include random effects of clinic (b_j), patient (c_{ijk}) and residual (e_{ijkl}) at measurement times. We assume these are distributed $NID(0,\sigma^2_{center})$, $NID(0,\sigma^2_{patient})$, and $NID(0,\sigma^2_{error})$, respectively. Assume the population mean is $E(y_{ijkl}) = \mu$. Then the observation may be represented

$$y_{ijkl} = \mu + b_j + c_{ijk} + e_{ijkl}.$$

The variance in an observation due to sampling error is

$$V(y_{ijkl}) = \sigma^2_{center} + \sigma^2_{patient} + \sigma^2_{error}.$$

Now consider the effects of administering the drugs. First, consider the effects on the mean. Let the $\mu_{il} = \mu + \alpha_i + \tau_l + (\alpha\tau)_{il}$ denote the population mean at visit *k* for patients administered drug *i,* where α_i, τ_l, and $(\alpha\tau)_{il}$ are the *fixed* effects due to drug, visit, and drug*visit interaction. Next, there is a possible *random* interaction effect $(\alpha b)_{ij}$ between clinic and drug. Assume $(\alpha b)_{ij}$ is distributed $NID(0,\sigma^2_{center*drug})$. Then an observation is represented

$$y_{ijkl} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + c_{ijk} + \tau_l + (\alpha\tau)_{il} + e_{ijkl}.$$

The mean and variance are

$$E(y_{ijkl}) = \mu + \alpha_i + \tau_l + (\alpha\tau)_{il}$$

and

$$V(y_{ijkl}) = \sigma^2_{center} + \sigma^2_{center*drug} + \sigma^2_{patient} + \sigma^2_{error}.$$

Here are statements for MIXED and GLM appropriate for this model:

```
proc mixed data=multcent;
class drug patient inv visit;
model si_hr=drug visit drug*visit / ddfm=kr htype=1,2,3;
random inv drug*inv patient(drug*inv);
lsmeans drug / pdiff cl;
run;

proc glm data=multcent;
class drug patient inv visit;
model si_hr=inv drug drug*inv patient(drug*inv) visit drug*visit / ss1 ss2
ss3;
random inv drug*inv patient(drug*inv)/test;
lsmeans drug / pdiff;
run;
```

Notice the similarities and differences of code between MIXED and GLM:

1. CLASS statements are the same.
2. MODEL statement contains only fixed effects in MIXED, but all effects in GLM. Options are different.
3. RANDOM statements are basically the same, but GLM has the test option.
4. LSMEANS statements are basically the same, but MIXED has cl option.

The purpose of this paper is to chronicle the evolution of linear models in SAS® from the perspective of an outsider

COMPARISON OF MIXED AND GLM OUTPUT

Results from the MODEL and RANDOM statements.

REML estimates of variance components from MIXED:

Covariance Parameter Estimates	
Cov Parm	**Estimate**
inv	0.4369
drug*inv	4.7748
patient(drug*inv)	30.7752
Residual	62.7038

Tests of fixed effects from MIXED:

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
drug	1	29.6	4.24	0.0485
visit	4	946	0.36	0.8399
drug*visit	4	946	1.46	0.2135

Analysis of variance from GLM, which uses residual error mean squares for the F-tests:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
inv	52	12852.39795	247.16150	3.95	<.0001
drug	1	980.64667	980.64667	15.68	<.0001
drug*inv	43	8382.44479	194.94058	3.12	<.0001
patient(drug*inv)	214	37518.67326	175.32090	2.80	<.0001
visit	4	65.37280	16.34320	0.26	0.9028
drug*visit	4	409.79437	102.44859	1.64	0.1626

Next is the table of expected mean squares from GLM:

Source	Type III Expected Mean Square
inv	Var(Error) + 3.6493 Var(patient(drug*inv)) + 9.7134 Var(drug*inv) + 18.739 Var(inv)
drug	Var(Error) + 3.1135 Var(patient(drug*inv)) + 7.1417 Var(drug*inv) + Q(drug,drug*visit)
drug*inv	Var(Error) + 3.6142 Var(patient(drug*inv)) + 10.844 Var(drug*inv)
patient(drug*inv)	Var(Error) + 3.8199 Var(patient(drug*inv))
visit	Var(Error) + Q(visit,drug*visit)
drug*visit	Var(Error) + Q(drug*visit)

Test of fixed effects of drug from the GLM expected mean squares (note error DF):

	Source	DF	Type III SS	Mean Square	F Value	Pr > F
*	drug	1	980.646667	980.646667	5.72	0.0192
	Error	75.568	12951	171.387334		

Error: 0.6586*MS(drug*inv) + 0.1919*MS(patient(drug*inv)) + 0.1495*MS(Error)

*** This test assumes one or more other fixed effects are zero.**

Tests of other effects from GLM based on expected mean squares:

	Source	DF	Type III SS	Mean Square	F Value	Pr > F
	patient(drug*inv)	214	37519	175.320903	2.80	<.0001
*	visit	4	65.372799	16.343200	0.26	0.9028
	drug*visit	4	409.794373	102.448593	1.64	0.1626
	Error: MS(Error)	903	56490	62.557606		

*** This test assumes one or more other fixed effects are zero.**

Non-significant effects of visit and drug*visit interaction from both MIXED and GLM indicate that comparisons of overall drug means are justified.

Look at the least squares means from MIXED:

Least Squares Means

Effect	drug	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
drug	1	77.1892	0.7481	57.3	103.19	<.0001	0.05	75.6914	78.6869
drug	4	75.0811	0.7184	57.8	104.51	<.0001	0.05	73.6429	76.5192

Now the table of difference between the drugs from MIXED:

Differences of Least Squares Means

Effect	drug	_drug	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
drug	1	4	2.1081	1.0243	29.6	2.06	0.0485	0.05	0.01490	4.2013

Here's what we get from GLM:

drug	si_hr LSMEAN
1	Non-est
4	Non-est

This illustrates the dreaded non-est message. If comes about in this example because there are no patients on drug 1 in four of the clinics, and likewise no patients on drug 4 in four different clinics. GLM tries to average LSMEANS across all combinations of drug and clinic, but cannot do so because of the empty cells.

This does not happen in MIXED because clinic is a random factor. Do not get too comfortable and think the non-est problem will never occur in MIXED. Remember that non-estimability is an issure relative to two of more fixed factors. For example, drug and visit fixed factors. If there were no patients on drug 1 with measurements at time 5, then the LSMEANS for drug 1 and time 5 would be non-estimable; as well, of course, the LSMEAN for the combination of drug 1 and time 5.

CONCLUSION

SAS contains more than 50 statistical procedures. Of these, there are about a dozen mainstream procedures based more or less on linear models that probably account for 90% (just a guess) of the GNDA (gross national data analyses). One line of these contains GLM, MIXED and GLIMMIX, each representing a quantum increase in capability above the previous. My belief is that it will continue as computing power increases. Perhaps a linear models procedure of the future will handle "object" data. The "observation" will not be a number or character, but rather an assembly information such as a geographical image or an assembly of genetic material. Hopefully, SAS Institute will stay in the business of producing high-quality procedures.

REFERENCES

- Belsley, D.A., Kuh, E. and Welsch, R.E. (1980), ***Regression diagnostics: Identifying Influential Data and Sources of Collinearity***. New York: John Wiley and Sons.
- Fai, A.H.T. and Cornellius, P.L. (1996), "Approximate ***F***-tests of Multiple Degree of Freedom Hypotheses in Generalized Least Squares Analyses of Unbalanced Split-plot Experiments," ***J. of Statistical Computation and Simulation***, 54, 363-378.
- Geisbrecht, F.G. (1989), "A General Structure for the Class of Mixed Linear Models," ***Applications of Mixed Models in Agriculture and Related Disciplines***, Southern Cooperative Series Bulletin No. 343, Louisiana Agricultural Experiment Station, Baton Rouge, 183-201.
- Geisbrecht, F.G. and Burns, J.C. (1985), "Two-Stage Anallysis Bases on a Mixed Model: Large-sample Asymptotic Theory and Small-sample Simulation Results," ***Biometrics***, 42, 477-486.
- Goodnight, J.H. (1979), "A Tutorial on the Sweep Operator," ***The American Statistician***, 33, 149-158.
- Grizzle,
- Harvey, W.H. (1975), "Least-Squares Analysis of Data with Unequal Subclass Numbers," Bulletin ARS H-4, USDA.
- Harville, D.A., and Jeske, D.R. (1992), "Mean Squared Error of Estimation or Prediction Under a General Linear Model," ***J. of American Statistical Association***, 87, 724-731.
- Henderson, C.R. (1984), ***Applications of Linear Models in Animal Breeding***, University of Guelph.
- Hocking, R.R., H.,
- Huynh, H., and Feldt, L.S. (1970), "Conditions Under Which Mean Square Rations in Repeated Measurements Designs Have Exact ***F***-Distributions, ***J. of American Statistical Association***, 65, 1582-1589.
- Kackar, R.H., and Harville, D.A. (1984), "Approximations for Standard Errors of Estimation of Fixed and Random Effects in Mixed Linear Models," ***J. of American Statistical Association***, 79, 853-862.
- Kenward, M.G, and Roger, J.H. (1997), "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood," ***Biometrics***, 53, 983-997.

- Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., and Schagenberger, O. (2006), ***SAS for Mixed Models***, 2nd Edition, Cary, NC: SAS Institute Inc.
- Littell, R.C., Stroup, W.W., and Freund, R.J. (2005), ***SAS for Linear Models***, 4th Edition, Cary, NC: SAS Institute Inc.
- McLean, R.A., and Sanders, W.L. (1988), "Approximating Degrees of Freedom for Standard Errors in Mixed Linear Models," ***Proceedings of the Statistical Computing Section***, American Statistical Association, New Orleans,50-59.
- McLean, R.A., and Sanders, W.L., and Stroup, W.W. (1991), " ***The American Statistician***, 45, 54-64.
- Nelder, J.A., and Wedderburn, R.W.M., "Generalized Linear Models," ***J. of the Royal Statistical Society,Series A***, 135, 370-384.
- Patterson, H.D., and Thompson, R. (1974), "Recovery of Intra-block Information When Block Sizes Are Unequal," ***Biometrika***, 58, 545-554.
- Prasad, N.G.N., and Rao, J.N.K., (1990), "The Estimation of Mean Squared Error of Small-Area Estimators," ***J. of American Statistical Association***, 85, 163-171.
- Searle, S.R. (2006), ***Linear Models for Unbalanced Data***, New York: John Wiley and Sons.
- Snedecor, G.W., and Cochran, W.G. (1967), ***Statistical Methods***, 6th Edition, Ames, IA, Iowa State University Press.
- Speed, F.M., Hocking, R.R., and Hackney, O.P. (1978), "Methods of Analysis of Linear Models with Unbalanced Data,"
- Steel, R.G.B., and Torrie, J.H. (1960), ***Principles and Procedures of Statistics***, New York: McGraw-Hill Book Co.
- Stroup, W.W. (1989), "Predictable Functions and Prediction Space in the Mixed Model Procedure," ***Applications of Mixed Models in Agriculture and Related Disciplines***, Southern Cooperative Series Bulletin No. 343, Louisiana Agricultural Experiment Station, Baton Rouge, 39-48.
- Swallow, W.H., and Monahan, J.F. (1984), "Monte Carlo Comparison of ANOVA, MIVQUE, REML, and ML Estimators of Variance Components," ***Technometrics***, 28,47-57.

ACKNOWLEDGMENTS

I wish to thank Debbie Buck, Jennifer Waller, and Rachael Biel for the opportunity to present this paper, and for their patience and assistance in making arrangements. Also, I wish to thank my friend and co-author Walt Stroup for all the conversations about statistics and life.

CONTACT INFORMATION

Name: Ramon C. Littell
Enterprise: Info Tech Inc
Address: 5700 SW 34th St.
City, State ZIP: 32608
E-mail: littell@ufl.edu