Paper 258-2009

Comparison of Features and Applications of four Linear Models Procedures

Larry W. Douglass, University of Maryland, College Park, MD

Abstract

Development of SAS® linear models procedures over the past several years has led to a number of easily accessible methodological statistical advances for experimental data analysis. The original linear models program, GLM, was a fixed model procedure for analysis of normally distributed data with homogeneous variances. The GENMOD procedure extended the fixed linear model analysis to a number of non-normal distributions. With the use of GEE, GENMOD was able to address correlated repeated measures data. The MIXED procedure permitted the user to model both fixed and random effects for normally distributed variables. Because modeling of random effects permits multiple residual error terms, it is frequently possible to model heterogeneous residual variances. The most recent linear models procedure, GLIMMIX, has the capabilities of GLM, GENMOD and MIXED in one procedure. This presentation looks at the unique features and appropriate applications of each of these linear model procedures.

Introduction

The development of SAS® linear models procedures over the past several years has led to a number of easily accessible methodological statistical advances for experimental data analysis. GLM, GENMOD, MIXED and GLIMMIX are linear models procedures developed for the analysis of experimental data from designed experiments. GLM and GENMOD were developed to model fixed effects, while MIXED and GLIMMIX were developed to model both fixed and random effects. GLM and MIXED are limited to normally distributed data, while GENMOD and GLIMMIX were developed to analyze data from the exponential family of distributions. Because GLIMMIX was developed to analyze both fixed and random effects and data from the exponential family of distributions, which includes the normal distribution one would expect GLIMMIX to appropriately analyze data that could be appropriately analyzed by any of the other three linear model procedures.

This paper will include a brief description of fixed versus mixed models and identify some of the common members of the exponential family of distributions. A discussion of the four linear models procedures, with examples to illustrate their limitations and features that make them useful for certain analysis situations. The remaining portion of the paper will present a number of examples illustrating data analyses using GLIMMIX. The objective of this paper is to make users of these SAS linear models procedures aware of their limitations, as well as, their appropriate use for the analysis of a broad range of experimental data.

Overview of Linear Models

The following figure illustrates the relationship between these linear model procedures.

	Normal distributions	Exponential family of distributions
Fixed models	GLM	GENMOD
Mixed models	MIXED	GLIMMIX

Factors in an experiment and in a linear model can be partitioned into two categories. Treatment structure which consist of factor levels that the researcher wishes to examine and/or compare. These include the usual treatment factors such as drugs, drug levels, levels of dietary nutrients, type or level of exercise, etc. and are known as fixed effects. Design structure which consists of factors that identify how experimental units were grouped to create more homogeneous sets of experimental units. These include factors such as blocks, days, locations, lab technicians, pen, etc. and are known as random effects.

A simple formulation of fixed model would be:

Y = {treatment structure components} + {error structure}

A similar formulation of a mixed model would contain one addition set of components:

Y = {treatment structure components} + {design structure components} + {error structure}

In the mixed model error terms are normally defined as interactions of treatment structure and design structure factors.

A factor is fixed when the levels of the factor are selected by researcher. A fixed factor is one that is repeatable. Because the levels of a fixed effect are selected by the researcher exactly the same factor levels could be repeated. That is, if you or other scientists repeat your experiment, they would be estimating the same differences among treatment means, the same covariate regression coefficients or the same differences among regression coefficients. For a random effect the levels are randomly selected from a population of possible levels. The levels of a random effect could not be repeated, because another random sample of levels would result in different levels of the factor. That is, if you or other scientists repeat your study you would not (probably could not) estimate the same effects, but could provide an estimate the same variance of the random factor. The distinction is that a repetition of a random factor would include different levels of the random factor, while repetition of a fixed factor would include exactly the same levels of the factor.

Suppose that an experiment is conducted at three locations (clinics). Locations should be modeled as fixed if a repetition of the experiment would be conducted at the same locations. The statistical inference would be restricted to those three locations. Locations should be modeled as random if a repetition of the experiment would result in a different set of locations. The statistical inference would be to the population the locations would reasonably represent. For example, if the experiment includes three clinics from Atlanta and a repetition of the experiment would be conducted on other Atlanta clinics, than model the location as random and limit the inference to Atlanta clinics.

GLM and MIXED are known as general linear models procedures. General linear models are a specific case of a larger class of models known as generalized linear models. GENMOD and GLIMMIX are generalized linear models procedures. Generalized linear models allow us to analyze data where the distribution is a member of the exponential family of distributions. The normal distribution is a member of the exponential family of distributions. Generalized linear models include error distributions for modeling normal, binary, binomial, negative binomial, Poisson, multinomial and several other distributions. These models allow the treatment means to be modeled by selection of an appropriate link function and the error probability distribution.

Appropriate Applications and Limitations

GLM was available in the earliest versions of SAS and was for years the mainstay of linear models analysis of experimental data. GLM, an ordinary least squares procedure, was developed for balance or unbalanced fixed model analysis of variance. The assumptions of normality, homogeneity of variances and independence were required. The only design structure that could be reasonably characterized as a fixed effects model is a completely randomized design structure with equal treatment variability and uncorrelated errors. If this is your data than GLM is appropriate.

What are some of the problems commonly encountered using GLM?

GLM was frequently used for mixed model analysis. SAS made an effort to improve GLM capabilities for mixed model analysis with the addition of various options and statements. The E=option on the TEST, CONTRAST and LSMEANS statements and the TEST option on the RANDOM statement are examples. At the time, these attempts resulted in significant improvements in our ability to analyze mixed models using available statistical software packages. In the mid 80's a book by Milliken and Johnson (The Analysis of Messy), pointed out the problems associated with using GLM for mixed model analysis by illustrating ways to correct GLM's output when analyzing mixed models.

Even for the simplest mixed model analysis, a randomized complete block design, GLM can not compute the appropriate standard errors of the mean. The standard errors of differences, t and F ratios are correct. So at least our tests of hypotheses are correct, but the SEM does not include the block variance as it should when blocks are random.

Standard errors should reflect the variation in the statistic that would be expected in repetitions of the study. Therefore, variances associated with random effects, not just the residual variance, may contribute to the magnitude of the standard errors. Because GLM was developed as a fixed model program, GLM will not correctly compute standard errors that involve more than one random source of variation. Thus, except for the simplest designs (fixed models) some GLM standard errors are incorrect. The E= option is limited to the specification of one, and only one, random source of variation, while some tests require the combination of two or more random sources of variation. Although the TEST option of the RANDOM statement will combine variation from multiple sources of variation, it is a post model fitting fix-up that is applied only to the F tests in the analysis of variance.

F ratios and standard errors are presented in table 1 to illustrate which common statistics are incorrectly computed for a relatively simple mixed model analysis. The design was a completely randomized split plot design with four dietary treatments assigned to subjects and data were recorded at two times (pre and post treatment). Incorrect values are underlined. The values in the first column are those from a completely balanced experiment (no missing data). Although the ANOVA tests of significance are correct, the contrasts testing the difference between diets within a time are incorrect. Tests of significance using the PDIFF option of the LSMEANS statement generates the same incorrect tests (not shown). In addition, the standard error of the differences, generated by the ESTIMATE statement, are incorrect for both the differences between diet main effect means and between diet means within a time.

To generate the last two columns four observation were deleted from the data set. The column labeled 2WU, indicating that two whole unit were deleted (both pre and post observations). The incorrect statistics in this case are the same ones that were incorrect for the completely balance data set. However, if four observations (pre or post) are deleted each from a different whole, column labeled 4SU, all of the F ratios and standard errors are incorrect.

Table 1: Selected statistics from GLM analysis of a completely randomized split plot design with incorrect values underlined.

	F ratios		
ANOVA	Complete[a]	2WU[b]	4SU[c]
D	18.16	15.42	19.36
T	271.35	259.30	152.52
D*T	6.54	8.01	5.71
	F ratios		
CONTRAST	Complete	2WU	4SU
d1-d2	6.35	4.16	5.05
t0-t1	271.35	259.30	152.52
d1-d2 @ t1	50.26	49.37	22.69
t0-t1 @ d1	56.35	61.71	28.47
	Standard errors of the difference		
ESTIMATE	Complete	2WU	4SU
d1-d2	.565	.604	.893
t0-t1	.400	.427	.557
d1-d2 @ t1	.800	.854	1.575
t0-t1 @ d1	.800	.764	1.031
	Standard errors of the mean		
LSMEANS	Complete	2WU	4SU
d1	1.216	1.322	1.349
t0	.283	.302	.298
d1t0	.565	.540	.595

[a] No missing observations, completely balanced designed.
[b] Four missing observations, both observations (pre and post) on two whole units.
[c] Four missing observations, one each from 4 four different whole units.

Non-estimability of least squares means occurs when a treatment combination(s) is missing in a factorial treatment structure. Suppose that in a 2x3 factorial, the treatment combination representing the last level of each factor (a2b3) is missing. Since a main effect mean for one factor is obtained by averaging across all levels of the other factor, the main effect means for last level of each factor (a2 and b3) are by definition non-estimable. GLM or MIXED would correctly report in place of the least squares means the message 'NON-EST' and a dot (.) for corresponding standard errors and tests significance. Because GLM considers all factors as fixed when fitting the model, the same result occurs in GLM when the interaction of a random and fixed effect, with a missing combination, is included in the model to define an error term. MIXED will in this case estimate the main effect least squares means for all levels of the fixed effect.

Another example incomplete mixed model output with GLM would be the addition of a pretreatment covariate to the model, where the covariate was measured once on each subject prior to the start of the experiment where it is also necessary to include the animal effect in the model. In this case the covariate is confounded with the animal effect and although the animal and treatment sources are adjusted for the covariate, no tests of significance about the covariate are provided by GLM. MIXED will provide appropriate test of significance about the covariate effect.

Correlated Data (Temporal or Spatial): A commonly taught assumption of ANOVA is the 'Independence of model residuals'. Although GLM has a repeated measures statement that may provide correct results in some cases, in most cases it is either inefficient or not the most appropriate analysis. The GLM repeated measures tests are based on assumptions that are frequently not true. However, contrasts are available which, if appropriate for your data, are correct, but conservative. MIXED allows you to model the variances and covariances (correlations) among temporally or spatially correlated data and makes use of the variances and covariances to estimate the standard errors and to test hypotheses. Repeated measures could be analyzed in GLM as a multivariate analysis of variance using the MANOVA statement. However this analysis would be conservative if a simpler covariance structure was appropriate.

Heterogeneous variances: The residual variance in GLM is pooled across all treatment groups, which leads to the assumption of homogeneity of variances among groups. MIXED contains features which allows the user to fit separate variances for different groups, such as different treatments and/or different time periods. For some analyses, partitioning the residual variance may be reasonable, thus in those cases the data may be analyzed without a transformation.

Negative estimates of variance components: In least squares analysis of variance F ratios less than one result from a negative estimate of a variance component. A negative component for a fixed source is of little concern for tests of fixed sources since small ratios simply indicate 'no effect'. However a negative variance component for a random source is of concern if the negative component becomes part the test of a fixed effect. The technique used in MIXED to fit the random effects portion of the model restricts the variance component estimates to be zero or positive values, although the estimates of variance are now biased.

GENMOD was developed to analyze data from the exponential family of distributions. The normal distribution is a member of the exponential family of distributions. Therefore, data appropriately analyzed with GLM could also be appropriately analyzed by GENMOD, although GLM may contain options that would make GLM's usage preferred. Generalized linear models include error distributions for modeling normal, binary, binomial, Poisson, negative binomial, Poisson, multinomial and several other distributions. These models allow the treatment means to be modeled by selection of an appropriate link function and the error probability distribution. GENMOD was developed for the analysis of balance or unbalanced data using fixed linear models procedures as described by Nelder and Wedderburn (1972).

Generalized estimating equations (Liang and Zeger, 1986) were added to GENMOD to deal with correlated data in generalized linear models. Although GEE's may provide an appropriate way to deal with correlated data in GENMOD, users may find techniques associated with mixed model techniques more satisfactory.

The only design structure that could be reasonably characterized as fixed effects model is a completely randomized design structure. If your data can reasonable be described by a member of the exponential family of distributions and the design structure is simple completely randomized experiment, then GENMOD is likely to provide an appropriate analysis. An additional limitation is that test statistics rely on asymptotic theory for some distribution. That is large samples may be needed.

The **MIXED** procedure permits the user to model both fixed and random effects for normally distributed variables. While the major limitation of the mixed procedure is the required normality probability distribution for the residual errors, it does contain many useful features for modeling normal data. Because modeling of random effects permits multiple residual error terms, it is frequently possible to model heterogeneous residual variances and the REPEATED statement allow us to model correlated data. Mixed also contains a rich set of influence statistics (similar to those in the REG procedure) that can assist the user in identifying observations or sets of observations that may be exerting undue influence on the results.

The features that make the MIXED procedure a valuable tool in linear models analysis are primarily controlled with the RANDOM and REPEATED statements. Unlike the random statement in GLM, the MIXED RANDOM statement allows the user to appropriately model the design structure of the experiment. With the REPEATED statement the user can model heterogeneous residual variances among treatments. If appropriate a different residual variance may be assigned to each treatment, rather than a single pooled residual variance as with GLM. The major use of the REPEATED statement is to model correlated residuals as in repeated measures analyses or for spatial correlated data. I will discuss repeated measures in the GLIMMIX section, but for mixed I have selected an example fitting heterogeneous variance using MIXED.

This experiment examines the growth curves of bacteria on different growth media. The experiment was a repeated measures design. However, since the researcher was interested in specific parameters of growth, non-linear growth curves were fit to each time series and the estimates of the growth parameters were analyzed using the MIXED procedure. In this case a repeated measures analysis is not necessary, since each time sequence is now represented by sample estimates that incorporate the time effects. The experiment is a CRD with 3 to 6 replications (rep) and six different growth media (medium). The dependent variable is the asymptotes (k) from the non-linear growth curve.

A one way analysis of variance, with 27 residual df, was run to generate residuals for examination of the ideal conditions of the analysis. Plots of the residuals indicated that some treatments had variances that were several times larger than those for the treatments with the smallest variance. The plots did not indication of a serious departure from normality. The following mixed model analysis was run to further examine the heterogeneous variability.

```
TITLE3 The dependent variable is K;
TITLE4 Fitting separate variances for each medium;
PROC MIXED DATA=est CL;
CLASS rep medium;
MODEL k  = medium / DDFM=KR;
REPEATED / GROUP=medium;
LSMEAN medium;
```

Growth Media Study
Bacterial Growth Curve Parameters
K are asymptotes for the non-linear growth curves

Obs	rep	medium	k
1	1	CO	8.82
2	1	CU	6.05
3	1	PE	8.59
4	1	RA	8.94
5	1	SO	8.28
6	1	SU	7.99
7	2	CO	8.99
8	2	PE	8.56
9	2	RA	7.75
10	2	SO	8.19
11	2	SU	7.82
12	3	CO	8.91
13	3	PE	8.73
14	3	RA	8.88
15	3	SO	8.14
16	3	SU	7.29
17	4	CO	9.45
18	4	CU	6.69
.	.	.	.
.	.	.	.
.	.	.	.
32	6	SO	8.22
33	6	SU	8.23

The following selected results were generated.

Growth Media Study
Bacterial Growth Curve Parameters
The dependent variable is K
Fitting separate variances for each medium

Class Level Information

Class	Levels	Values
rep	6	1 2 3 4 5 6
medium	6	CO CU PE RA SO SU

Covariance Parameter Estimates

Cov Parm	Group	Estimate	Alpha	Lower	Upper
Residual	medium CO	0.2479	0.05	0.09659	1.4912
Residual	medium CU	0.4560	0.05	0.1236	18.0124
Residual	medium PE	0.008377	0.05	0.003264	0.05039
Residual	medium RA	0.7111	0.05	0.2771	4.2774
Residual	medium SO	0.005787	0.05	0.002255	0.03481
Residual	medium SU	0.1261	0.05	0.04913	0.7584

These are the estimates of the random variances (Estimate) for the six media. The advantage of this analysis of variance is that the assumption of homogeneity of residual treatment variances is not required, since no variances are being pooled. However, normality is still an assumption. Remember that the assumption of normality is that each treatment mean was from a normally distributed population of treatment means. The disadvantage of using this model is that it results in smaller number of degrees of freedom for test of significance then when variances are partitioned (6.8) as compared to degrees of freedom for a single pooled variance (27) for the experiment.

Fit Statistics

-2 Res Log Likelihood	16.4
BIC (smaller is better)	37.4

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
5	31.44	<.0001

The chi-square value is the difference between the -2 Res Log Likelihood (-2LL) for the two models and the degrees of freedom is the difference in the numbers of random variance parameters fit in the two models. The results indicate that we should reject the null hypothesis ($p<.0001$) that the variances are homogeneous. This does not mean that six residual variances are the best way to fit the random variance, but that it is a significantly better fit than a single pooled estimate of experimental variance.

There are two other candidate models that could be considered for these data. 1) Two variance groups: PE and SO pooled with one experiment variance and the other four media pooled to form a second variance. 2) Three variance groups of two media each, forming a small, an intermediate and a large variance group. To illustrate the fitting process I will present the two variance grouping.

```
DATA vargrp;
SET est;
IF medium IN('PE','SO') THEN vargrp=1;
IF medium IN('CO','CU','RA','SU') THEN vargrp=2;

PROC MIXED DATA=vargrp CL;
CLASS rep medium vargrp;
MODEL k  = medium / DDFM=KR OUTP=resids;
REPEATED / GROUP=vargrp;
LSMEAN medium / PDIFF;
```

Growth Media Study
Bacterial Growth Curve Parameters
The dependent variable is K
Fitting two separate experimental variance groups

Covariance Parameter Estimates

Cov Parm	Group	Estimate	Alpha	Lower	Upper
Residual	vargrp 1	0.007082	0.05	0.003457	0.02181
Residual	vargrp 2	0.3728	0.05	0.2099	0.8378

These are the estimates of variances for the two media groups. The variance for group 2 is ~50 times greater than the variance for group 1.

Fit Statistics

-2 Res Log Likelihood	20.4
BIC (smaller is better)	27.4

The model fitting information provides statistics for evaluating the fit of the two parameter model as compared to the six parameter model.

	Number of parameters	
Criteria	Two	Six
-2LL	20.4	16.4
BIC	27.4	37.4
DF	19.2	6.8

AIC and BIC are penalized residual log likelihood (-2LL) statistics. -2LL is similar to an R-squared value in multiple regression. The more parameters added to the model the better the fit (smaller value). Just like R-square, the -2LL does not take into account the number of parameters (simplicity) in the model. BIC is more like an adjusted R-square, because their values have been increased (penalized) depending on the number of parameters in the model. In this case BIC supports the two as compared to the six variance model.

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	27.44	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
medium	5	19.2	22.14	<.0001

Least Squares Means

Effect	medium	Estimate	Standard Error	DF	t Value	Pr > |t|
medium	CO	9.3317	0.2493	17	37.44	<.0001
medium	CU	6.0267	0.3525	17	17.10	<.0001
medium	PE	8.5783	0.03436	10	249.70	<.0001
medium	RA	8.0400	0.2493	17	32.26	<.0001
medium	SO	8.2367	0.03436	10	239.75	<.0001
medium	SU	7.9400	0.2493	17	31.85	<.0001

Because **GLIMMIX** was developed to analyze both fixed and random effects and data from the exponential family of distributions, which includes the normal distribution one should expect GLIMMIX to appropriately analyze data that could be analyzed by any of the other three linear model procedures.

MIXED and GLIMMIX both have a full suite of variance-covariance matrices for analysis of repeated measures experiments. The example presented below was from an experiment examining the effects flyicides on numbers of flies on cattle. The treatments are control (0) and three flyicides. Fly counts associated with each animal were taken at 2, 4 and 6 months post treatments. Because the dependent variable (flies) is a count, Poisson repeated measures should be considered.

```
data flyicide;
input animal$ treatment month flies;
datalines;
    a           0          2        13
    a           0          4        10
    a           0          6        13
    b           2          2        10
    b           2          4         8
    b           2          6         8
    c           3          2        10
    .           .          .         .
    .           .          .         .
    .           .          .         .

    p           2          6         6
run;
```

One choice for analysis would be to use a normalizing transform and use MIXED to analyze the transform data. Using this approach would require the following MIXED code to fit the data to an unstructured variance-covariance matrix.

```
PROC MIXED DATA=flyicide;
CLASS animal treatment month;
MODEL transform_flies = treatment month treatment*month / DDFM=KR;
REPEATED month / SUBJECT=animal(treatment) TYPE=un;
LSMEANS treatment month treatment*month;
```

GLIMMIX does not have a REPEATED statement. The functionality of the repeated statement has been incorporated into the RANDOM statement of GLIMMIX. To analyze fly counts as Poisson distributed repeated measures, the GLIMMIX procedure would require the following code.

```
PROC GLIMMIX DATA=flyicide;
CLASS animal treatment month;
MODEL flies = treatment month treatment*month / DDFM=KR DIST=POISSON LINK=LOG;
RANDOM month / RESIDUAL SUBJECT=animal(treatment) TYPE=un;
LSMEANS treatment month treatment*month / ILINK;
```

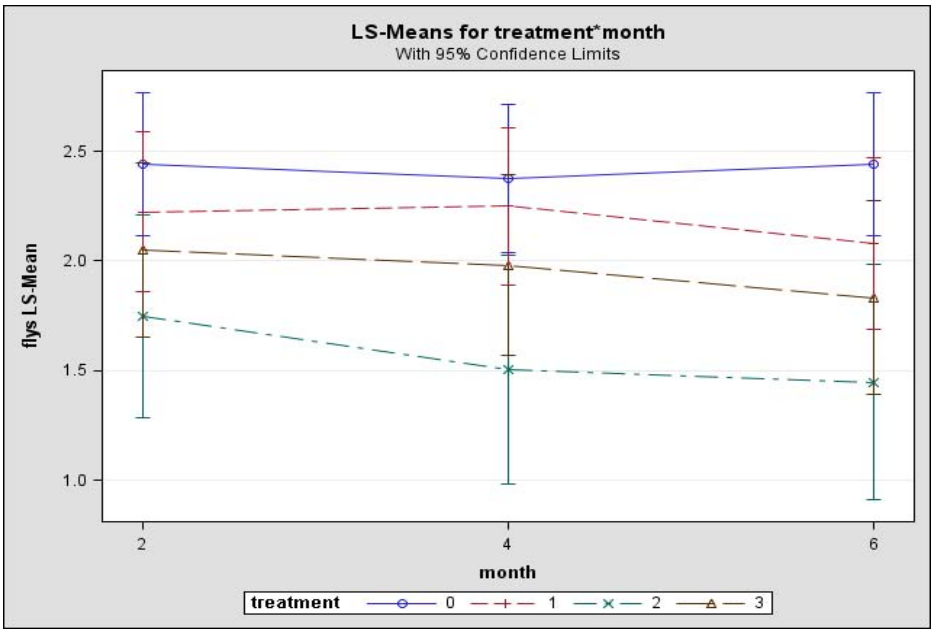
The resulting unstructured variance-covariance matrix.

Repeated	Repeated Measure		
Measure	month2	month4	month6
month2	1.0493	0.6321	0.9688
month4		1.1029	0.8778
month6			1.1274

The covariance matrix has been scaled such that if the data were exactly Poisson then the diagonal values would be 1.0. The diagonals are estimates of the overdispersion parameter. For the Poisson distribution, Variance=Mean, for overdispered Poisson the Variance=Overdispersion x Mean. The covariances are also scaled and reflect the relative magnitude of the correlation among residuals.

The results of the unstructured variance-covariance matrix suggest that compound symmetry might be another possibility. The following are results of the CS analysis. I have also included a number of options to illustrate some of the mean comparison features available in GLIMMIX. I have included ODS graphics statements to generate a two-way plot of the treatment means with their confidence limits that would be especially useful for examination of interactions.

```
ODS HTML;
ODS GRAPHICS ON;
ODS SELECT MEANPLOT;
TITLE2 Compound Symmetry;
PROC GLIMMIX DATA=flyicide;
CLASS animal treatment month;
MODEL flys = treatment month treatment*month/ DDFM=KR DIST=POISSON LINK=LOG;
RANDOM month / RESIDUAL SUBJECT=animal(treatment) TYPE=CS;
LSMEANS treatment*month / plot=MEANPLOT (SLICEBY=treatment JOIN CL);
```



A second lsmeans statement includes the ILINK option that request that the inverse link be applied to provide a mean on the original scale along with SEM using the delta method. The LINES option assigns letters to the means identifying significant differences among the least squares means. The SLICEDIFF=month computes the tests of simple effects between treatments for each month (18 comparisons).

```
LSMEANS treatment*month / ILINK LINES SLICEDIFF=month;
```

The high lighted columns are the result of the ILINK option and are on the original count scale.

treatment*month Least Squares Means

treatment	month	Estimate	Standard Error	DF	t Value	Pr > |t|	Mean	Standard Error Mean
0	2	2.4423	0.1542	16.8	15.84	<.0001	11.5000	1.7728
0	4	2.3749	0.1594	16.8	14.89	<.0001	10.7500	1.7140
0	6	2.4423	0.1542	16.8	15.84	<.0001	11.5000	1.7728
1	2	2.2246	0.1719	16.8	12.94	<.0001	9.2500	1.5900
1	4	2.2513	0.1696	16.8	13.27	<.0001	9.5000	1.6113
1	6	2.0794	0.1848	16.8	11.25	<.0001	8.0000	1.4786
2	2	1.7492	0.2180	16.8	8.02	<.0001	5.7500	1.2536
2	4	1.5041	0.2464	16.8	6.10	<.0001	4.5000	1.1090
2	6	1.4469	0.2536	16.8	5.71	<.0001	4.2500	1.0777
3	2	2.0477	0.1878	16.8	10.90	<.0001	7.7500	1.4554
3	4	1.9810	0.1942	16.8	10.20	<.0001	7.2500	1.4076
3	6	1.8326	0.2091	16.8	8.76	<.0001	6.2500	1.3069

The letters indicating significance among the treatment means are the result of the LINES option

T Grouping for treatment*month Least Squares Means (Alpha=0.05)

LS-means with the same letter are not significantly different.

	treatment	month	Estimate				
	0	2	2.4423			A	
	0	6	2.4423			A	
	0	4	2.3749	B		A	
	1	4	2.2513	B		A	C
	1	2	2.2246	B		A	C
	1	6	2.0794	B	D	A	C
	3	2	2.0477	B	D	A	C
	3	4	1.9810	B	D	A	C
	3	6	1.8326	B	D		C
	2	2	1.7492		D		C
	2	4	1.5041		D		
	2	6	1.4469		D		

Tests of simple effects resulting from the SLICEDIFF= option of the LSMEANS statement.

Simple Effect Comparisons of treatment*month Least Squares Means By month

Simple Effect Level	treatment	_treatment	Estimate	Standard Error	DF	t Value	Pr > |t|
month 2	0	1	0.2177	0.2309	16.8	0.94	0.3591
month 2	0	2	0.6931	0.2670	16.8	2.60	0.0190
month 2	0	3	0.3947	0.2430	16.8	1.62	0.1229
month 2	1	2	0.4754	0.2776	16.8	1.71	0.1052
month 2	1	3	0.1769	0.2546	16.8	0.69	0.4966
month 2	2	3	-0.2985	0.2877	16.8	-1.04	0.3143
month 4	0	1	0.1236	0.2328	16.8	0.53	0.6024
month 4	0	2	0.8708	0.2935	16.8	2.97	0.0087
.							
.							
.							
month 6	1	3	0.2469	0.2791	16.8	0.88	0.3889
month 6	2	3	-0.3857	0.3287	16.8	-1.17	0.2570

```
LSMESTIMATE treatment*month 'Month=6, trt 1 vs control' 0 0 1  0 0 -1  0 0 0  0 0 0;

ESTIMATE 'Month=6, trt 1 vs control' treatment 1 -1 0 0;
                                     treatment*month 0 0 1  0 0 -1  0 0 0  0 0 0;
      RUN;
      ODS GRAPHICS OFF;
      ODS HTML CLOSE;
```

The following was generated using the new LSMESTIMATE statement. The same results could have been obtained the ESTIMATE statement. The LSMESTIMATE simplifies the writing of contrasts because the coefficients are applied to the LSMEANS rather that to the estimate of effects as with the ESTIMATE statement.

Least Squares Means Estimates

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
treatment*month	Month=6, trt 1 vs control	0.3629	0.2407	16.8	1.51	0.1502

The following output is from the compound symmetry repeated measures analysis and is for comparison to the unstructured repeated measures analysis to determine how best to model these repeated measures data.

Using MIXED for Repeated Measures Analysis
Compound Symmetry

The GLIMMIX Procedure
Fit Statistics

-2 Res Log Pseudo-Likelihood	26.09
Generalized Chi-Square	9.61
Gener. Chi-Square / DF	0.27

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
CS	animal(treatment)	0.8262	0.3745
Residual		0.2669	0.07706

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
treatment	3	12	3.94	0.0362
month	2	24.6	2.94	0.0714
treatment*month	6	24.51	0.81	0.5704

The second GLIMMIX example examines a set of binary observations. Data were collected on a single herd of ~1000 dairy animals. The variable of interest was evidence of infection with Neospora coded 1 if infected and 0 if not infected. Neospora is a serious problem in infected herds because infected animals abort their calves. Of interest was the infection rate of neospora as a function of gender, location and differences in age within location.

NEOSPORA DATA

Obs	ID	Inf	Elisa	Age_mo	Location	Sex	adj_age
1	1015HNB	1	474	2.0	1	F	0.056
2	1018HNB	0	474	0.8	1	F	-1.144
3	1041HNB	0	948	1.1	1	F	-0.844
4	1067HNB	0	441	1.3	1	F	-0.644
5	1086HNB	0	640	2.3	1	F	0.356
6	1101HNB	0	731	2.6	1	F	0.656
7	1115HNB	0	1707	1.5	1	F	-0.444
8	1117HNB	0	492	2.8	1	F	0.856
9	1126HNB	0	200	2.5	1	F	0.556
.							
.							
.							

For the first example the model will include location and gender effects in a 2x2 factorial. The data are binary and logistic analysis will be used to model the response variable. The error distribution will be defined as binomial and the link as logit. The GLIMMIX program was as follows:

```
proc glimmix;
where location in(1,2);
class location sex;
model Inf = location sex location*sex
      / dist=bin link=logit;
lsmeans location sex location*sex / ilink;
run;
```

Now let's look at selected results of the analysis.

NEOSPORA DATA
The GLIMMIX Procedure

Model Information

Response Variable	Inf
Response Distribution	Binary
Link Function	Logit

Class Level Information

Class	Levels	Values
Location	2	1 2
Sex	2	F M

Fit Statistics

-2 Log Likelihood	365.62
Pearson Chi-Square / DF	1.01

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Location	1	342	2.56	0.1104
Sex	1	342	1.64	0.2012
Location*Sex	1	342	0.02	0.8768

Location Least Squares Means

Location	Estimate	Standard Error	DF	t Value	Pr > |t|	Mean	Standard Error Mean
1	-1.5208	0.2545	342	-5.98	<.0001	0.1793	0.03746
2	-1.0348	0.1656	342	-6.25	<.0001	0.2622	0.03204

Sex Least Squares Means

Sex	Estimate	Standard Error	DF	t Value	Pr > |t|	Mean	Standard Error Mean
F	-1.0833	0.2036	342	-5.32	<.0001	0.2529	0.03847
M	-1.4722	0.2253	342	-6.54	<.0001	0.1866	0.03419

Location*Sex Least Squares Means

Sex	Location	Estimate	Standard Error	DF	t Value	Pr > |t|	Mean	Standard Error Mean
F	1	-1.3499	0.2999	342	-4.50	<.0001	0.2059	0.04903
M	1	-1.6917	0.4113	342	-4.11	<.0001	0.1556	0.05403
F	2	-0.8168	0.2755	342	-2.96	0.0032	0.3065	0.05855
M	2	-1.2528	0.1839	342	-6.81	<.0001	0.2222	0.03179

For the second example the model will include location a class variable, age linear, age quadratic and the location by age linear interaction. Age is adjusted to a mean of zero for each location. In the GLIMMIX the error distribution will be defined as binomial and the link as logit. The GLIMMIX program was as follows:

```
proc glimmix;
where sex='F';
class location;
model Inf = location
            adj_age
            adj_age*adj_age
             location*adj_age
       /solution dist=bin link=logit;
lsmeans location / pdiff at mean ilink;
quit;
```

Now let's look at the results of the analysis.

Model Information

Response Variable	Inf
Response Distribution	Binary
Link Function	Logit

Class Level Information

Class	Levels	Values
Location	4	1 2 3 4

Fit Statistics

-2 Log Likelihood	959.88
Pearson Chi-Square / DF	1.01

Parameter Estimates

Effect	Location	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		-0.8890	0.1224	804	-7.26	<.0001
Location	1	-0.5112	0.3364	804	-1.52	0.1290
Location	2	0.06819	0.3025	804	0.23	0.8217
Location	3	0.4409	0.1851	804	2.38	0.0174
Location	4	0	.	.	.	.
adj_age		0.02409	0.006746	804	3.57	0.0004
adj_age*adj_age		-0.00027	0.000128	804	-2.08	0.0380
adj_age*Location	1	-0.4424	0.3287	804	-1.35	0.1788
adj_age*Location	2	-0.1116	0.1450	804	-0.77	0.4419
adj_age*Location	3	-0.07230	0.03260	804	-2.22	0.0269
adj_age*Location	4	0	.	.	.	.

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Location	3	804	3.45	0.0163
adj_age	1	804	2.16	0.1421
adj_age*adj_age	1	804	4.32	0.0380
adj_age*Location	3	804	2.42	0.0648

Location Least Squares Means

Location	adj_age	Estimate	Standard Error	DF	t Value	Pr > |t|	Mean	Standard Error Mean
1	0.00	-1.4002	0.3133	804	-4.47	<.0001	0.1978	0.04971
2	0.00	-0.8208	0.2768	804	-2.97	0.0031	0.3056	0.05873
3	0.00	-0.4481	0.1399	804	-3.20	0.0014	0.3898	0.03327
4	0.00	-0.8890	0.1224	804	-7.26	<.0001	0.2913	0.02528

Differences of Location Least Squares Means

Location	Location	adj_age	Estimate	Standard Error	DF	t Value	Pr > |t|
1	2	0.00	-0.5794	0.4181	804	-1.39	0.1661
1	3	0.00	-0.9521	0.3431	804	-2.77	0.0056
1	4	0.00	-0.5112	0.3364	804	-1.52	0.1290
2	3	0.00	-0.3727	0.3101	804	-1.20	0.2297
2	4	0.00	0.06819	0.3025	804	0.23	0.8217
3	4	0.00	0.4409	0.1851	804	2.38	0.0174

Previous in the paper I have addressed the analysis of response variables with only 2 possible outcomes (binary). What if our response variable has more than two possible outcomes (multinomial). The common approach is to use a chi-square contingency table analysis, which is equivalent to a one-way anova. But what if the treatment structure is a factorial? Glimmix provides an analysis for multi-way treatment structures when the response variable is multinomial.

The experiment was a survey examining the effects of air pollution (Low, High), job air quality (Good, Poor) and smoking status (Non, Ex, Current).

The response variable is Respiratory Disease Level (I, II, III or IV), with I being no symptoms and IV being the most severe.

The DIST= multinomial and LINK=cumlogit for an ordered multinomial response variable.

Chronic Respiratory Disease Data

****

Ordered Response Variable Definition

I = no symptoms,
II = cough or phlegm less than three months per year
III = cough or phlegm more than three months per year
IV = cough and phlegm plus shortness of breath more than three months per year

Air pollution status (Low, High)
Job air status (Good, Poor)
Smoking is smoking status (Non, Ex, Current)

****

There are 48 lines of data (combinations of air, job, smoking and respiratory disease level). Count represents the number of time each combination occurred in the survey. Count will be use as a frequency variable in the glimmix analysis. The total count (number of subjects in the survey is 2090. The data are presented below in two way tables.

Data frequencies for Smoking by RespiratoryLevel, Air=High and Job=Good

Smoking RespiratoryLevel

Frequency	I	II	III	IV
Current	77	48	39	51
Ex	39	11	4	2
Non	32	3	6	1

Data frequencies for Smoking by RespiratoryLevel, Air=High and Job=Poor

Smoking RespiratoryLevel

Frequency	I	II	III	IV
Current	184	65	33	36
Ex	67	8	4	3
Non	94	7	5	1

Data frequencies for Smoking by RespiratoryLevel, Air=Low and Job=Good

Smoking RespiratoryLevel

Frequency	I	II	III	IV
Current	94	48	46	60
Ex	38	12	4	4
Non	26	5	5	1

Data frequencies Smoking by RespiratoryLevel, Air=Low and Job=Poor

Smoking RespiratoryLevel

Frequency	I	II	III	IV
Current	307	102	83	68
Ex	167	19	5	3
Non	158	9	5	1

```
TITLE2 Multinomial distribution, Cumulative Logit;
PROC GLIMMIX DATA=Respiratory;
FREQ count;
CLASS Air Job Smoking;
MODEL RespiratoryLevel(DESCENDING) = Air Job  Smoking  Job*Smoking
       / SOLUTION DIST=MULTINOMIAL LINK=CUMLOGIT;
```

Chronic Respiratory Disease Data
Multinomial distribution, Cumulative Logit

Model Information

Response Variable	RespiratoryLevel
Response Distribution	Multinomial (ordered)
Link Function	Cumulative Logit

Class Level Information

Class	Levels	Values
Air	2	High Low
Job	2	Good Poor
Smoking	3	Current Ex Non

Response Profile

Ordered Value	Respiratory Level	Total Frequency
1	IV	231
2	III	239
3	II	337
4	I	1283

The GLIMMIX procedure is modeling the probabilities of levels of RespiratoryLevel having lower Ordered Values in the Response Profile table.

Parameter Estimates

Effect	Respiratory Level	Air	Job	Smoking	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	IV				-3.9848	0.2128	2081	-18.73	<.0001
Intercept	III				-3.0653	0.2062	2081	-14.86	<.0001
Intercept	II				-2.1868	0.2019	2081	-10.83	<.0001
Air		High			-0.04467	0.09380	2081	-0.48	0.6340
Air		Low			0	.	.	.	.
Job			Good		1.2091	0.3207	2081	3.77	0.0002
Job			Poor		0	.	.	.	.
Smoking				Current	1.9725	0.2095	2081	9.41	<.0001
Smoking				Ex	0.4642	0.2596	2081	1.79	0.0740
Smoking				Non	0	.	.	.	.
Job*Smoking			Good	Current	-0.4009	0.3377	2081	-1.19	0.2352
Job*Smoking			Good	Ex	-0.2587	0.4102	2081	-0.63	0.5283
Job*Smoking			Good	Non	0	.	.	.	.
Job*Smoking			Poor	Current	0	.	.	.	.
Job*Smoking			Poor	Ex	0	.	.	.	.
Job*Smoking			Poor	Non	0	.	.	.	.

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Air	1	2081	0.23	0.6340
Job	1	2081	48.40	<.0001
Smoking	2	2081	95.26	<.0001
Job*Smoking	2	2081	0.77	0.4616

The initial model was the full factorial model. I deleted the least significant high order sources first. In practice I would also delete the Job*Smoking, but elected to leave it in this example to show that interactions can be include in multinomial glimmix models.

The following are examples code for estimate statements was used compute the cumulative percentages for Job Status = Poor.

```
ESTIMATE 'Job=Poor IV'          INTERCEPT 6 0 0  Air 3 3  Job 6 0  Smoking 2 2 2
      job*smoking 2 2 2 0 0 0 / ILINK DIVISOR=6;
ESTIMATE 'Job=Poor III+IV'      INTERCEPT 0 6 0  Air 3 3  Job 6 0  Smoking 2 2 2
      job*smoking 2 2 2 0 0 0 / ILINK DIVISOR=6;
ESTIMATE 'Job=Poor II+III+IV'   INTERCEPT 0 0 6  Air 3 3  Job 6 0  Smoking 2 2 2
      job*smoking 2 2 2 0 0 0 / ILINK DIVISOR=6;
```

The cumulative percentages for Job Status Poor are in the Mean column in the following output.

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Mean	Standard Error Mean
Job=Poor IV	-2.2057	0.1235	2081	-17.86	<.0001	0.09924	0.01104
Job=Poor III+IV	-1.2862	0.1141	2081	-11.28	<.0001	0.2165	0.01935
Job=Poor II+III+IV	-0.4077	0.1105	2081	-3.69	0.0002	0.3995	0.02651

Summary of Respiratory Level percentages by Job Status generated using estimate statements

	Cumulative Percentage			
Job	I	II	III	IV
Poor	100	40	22	10
Good	100	20	9	4

	Percentage			
Job	I	II	III	IV
Poor	60	18	12	10
Good	80	11	5	4

Summary of applications and features

GLIMMIX can correctly perform most of the analyses that are can be appropriate analyzed using GLM, GENMOD and MIXED. In addition it is the most comprehensive of the linear models programs, since it can appropriately analyze both fixed and mixed models, and data from the exponential family of distributions. GLIMMIX's ods graphics output, statements and options for examining means are very useful additions. Also the use of log-linear models for mixed model analysis of multidirectional frequency data is a welcome tool. For some probability distributions, small samples may result in difficult fitting some models and a reduction in sensitivity. MIXED is limited to data with normally distributed errors. Influence statistics and the associated ods graphics output for examining assumptions of the analysis and identifying influential observations or set of observations, is not available in the other procedures. Both GLIMMIX and MIXED have a very comprehensive set of variance-covariance structures for analyzing correlated data, while GENMOD includes the GEE technique for correlated count data. GLM does include a MANOVA statement for fixed model multivariate analysis of variance. GLIMMIX and MIXED can conduct a form of multivariate analysis using the variance-covariance structures incorporate the correlation among response variables.

References

Liang, K. Y. and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. (2006), *SAS for Mixed Models*, Second Edition, Cary, NC: SAS Institute Inc.

Milliken, G. A. and Johnson, D. E. (1984), *Analysis of Messy Data, Volume I: Designed Experiments*, Belmont, CA: Lifetime Learning Publications.

Nelder, J. A. and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society A*, 135, 370–384.

Acknowledgements

Thanks to the members of NCCC-170 (USDA-CSREES Multi-State Research Committee) for their efforts in the development of education materials for mixed models.

CONTACT INFORMATION

For further information, you may contact:

Larry Douglass
Phone: 303-746-2162
Email: ldouglas@umd.edu