Paper 231-2011

Repeated Measures Analysis of Correlated Data with Multiple Responses using SAS®

Anbuchelvi Jeyabalasingham, Kamarajar University, Vavuniya, Sri Lanka
Anpalaki J Ragavan, Department of Mathematics and Statistics, University of Nevada, Reno, NV 89557, USA

ABSTRACT

When several measurements are taken on the same experimental unit, the measurements tend to be correlated with each other. When the measurements are responses to levels of an experimental factor of interest, such as time or treatment, the correlation can be taken into account by performing a repeated measures analysis of variance. In addition, if there are multiple responses, the analysis requires methods of multivariate analysis combined with tools of repeated measures, which is a challenging task and requires careful attention to several details such as distributional variation among responses, correlations among responses, subject specific trend in the treatment variable. In this research, repeated measures analysis of correlated data with multiple response variables that are a mixture of continuous, count, and binomial is explored. The common problems that arise when analyzing such data are addressed in detail in SAS. Data integrity testing, model selection and analysis of the data are presented in a step-by-step protocol with the SAS codes used in each step and the output shown. The advantages and disadvantages of using the SAS procedures, GLM, MIXED, and GLIMMIX are compared.

INTRODUCTION

When several measurements are taken on the same experimental unit (ex: person, animal, machine), the measurements tend to be correlated with each other. When the measurements represent qualitatively different things, such as weight, length, and width, this correlation is best taken into account by use of multivariate methods, such as multivariate analysis of variance. When the measurements are responses to levels of an experimental factor of interest, such as time, treatment, or dose, the correlation can be taken into account by performing a repeated measures analysis of variance. A popular repeated-measures design is the crossover study. A crossover study is a longitudinal study in which subjects receive a sequence of different treatments (or exposures). While crossover studies can be observational studies, many important crossover studies are controlled experiments. Repeated measures allow conducting an experiment when few participants are available. The repeated measure design reduces the variance of estimates of treatment-effects, allowing statistical inference to be made with fewer subjects. Repeated measures allows to conduct experiment more efficiently: Repeated measures designs allow many experiments to be completed more quickly, as only a few groups need to be trained to complete an entire experiment. For example, there are many experiments where each condition takes only a few minutes, whereas the training to complete the tasks take as much, if not more time. Repeated measures designs allow researchers to monitor how the participants change over the passage of time, both in the case of long-term situations like longitudinal studies and in the much shorter-term case of practice effects. It is possible that clustered data arise when multiple observations are collected on the same subject or experimental unit at different points in time or space, which leads to a special class of repeated measures, longitudinal, and spatial data, where multiple observations refer to different attributes. When there is clustered data structure, it is of interest to study the influence of clusters on the analysis rather than the influence of individual observations. A cluster comprises the repeated measurements for one or more subject. In this paper a repeated measures analysis on a data set with multiple responses, which included continuous, count, and binomial response variables (weight (Y1), continuous response), free food intake (Y2), binomial response, number of unplanned meals (Y3), count response), taken on 101 subjects over a period of twenty four weeks at four weeks interval (4^{th} , 8^{th} , 12^{th} and 24^{th} week). The data indicated clustered observations and correlations among the response variables. Each subject also received one of two treatments, the diet type (low carbohydrate (LC) or low fat (LF) diet). Three SAS procedures PROC MIXED, PROC GLIMMIX and PROC GLM were compared in performing the analysis of the data.

METHODS

ANALYSING DATA USING THE GLM PROCEDURE

One approach to analyze such data is to fit a triply multivariate repeated measures generalized linear model with PROC GLM with an IDENTITY statement along with the REPEATED statement. This differs from previous releases of PROC GLM, in which a MANOVA statement is used to perform a repeated measures analysis with multiple responses. Three responses, Y1, Y2, and Y3 are each measured four times for each subject (4th week, 8th week, 12th week and 24th week). Each subject received one of two treatments (low carbohydrate diet or a low fat diet). In PROC GLM, a REPEATED factor of type IDENTITY is used to identify the different responses and another REPEATED factor to identify the different measurement times. The repeated measures analysis includes multivariate tests for time and treatment main effects, as well as their interactions, across responses as produced by the following statements (SAS CODE 1).

SAS CODE 1

```
proc glm data=food;
class Treatment;
model Y1_4 Y1_8 Y1_12 Y1_24
        Y2_4 Y2_8 Y2_12 Y2_24
        Y3_4 Y3_8 Y3_12 Y3_24= Treatment / nouni;
repeated Response 3 identity, Time 4; run;
```

In the above statement RESPONSE is 1 for all the Y1 measurements, 2 for all the Y2 measurements and 3 for the Y3 measurements, while the four levels of TIME identify the Week4 Week8, Week12 and Week24 measurements within each response. A table of multivariate tests for within-subject effects, Response*Treatment tests for an overall treatment effect across the three responses, tables for Response*Time and Response*Treatment*Time test for time and the treatment-by-time interaction are output (see SAS output below, Table 1, Table 2, and Table 3).

Table 1

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no Response Effect
H = Type III SSCP Matrix for Response
E = Error SSCP Matrix
S=1 M=0.5 N=25.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.02682664	640.88	3	53	<.0001
Pillai's Trace	0.97317336	640.88	3	53	<.0001
Hotelling-Lawley Trace	36.27637396	640.88	3	53	<.0001
Roy's Greatest Root	36.27637396	640.88	3	53	<.0001

The first table (Table 1) displayed shows that the response effect is significant. In addition, the Response*Time interaction is significant, as shown in Table 2. However, the Response*Treatment interaction was not significant. Other tables are not shown due to space restriction. There is also a significant between subject effect for the treatment (p=0.0267) as shown in Table 3. In previous releases (before the IDENTITY transformation was introduced), in order to perform a triply repeated measures analysis, a MANOVA statement with a customized transformation matrix M had to be used. The MANOVA statement computed a univariate ANOVA for each transformed variable and provided a test for the overall main effect of time with the SUMMARY option (SAS CODE 2). The SUMMARY option in the MANOVA statement created an ANOVA table for each transformed variable as defined by the M matrix (differences of the 4th week, 8th week, 12th week, 24th week and so on). Results from SAS CODE 2 are not shown, but left for the reader to be obtained by running the SAS CODE given.

Table 2

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no Response*Time Effect
H = Type III SSCP Matrix for Response*Time
E = Error SSCP Matrix
S=1 M=3.5 N=22.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.15903454	27.61	9	47	<.0001
Pillai's Trace	0.84096546	27.61	9	47	<.0001
Hotelling-Lawley Trace	5.28794214	27.61	9	47	<.0001
Roy's Greatest Root	5.28794214	27.61	9	47	<.0001

Table 3

The GLM Procedure
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Treatment	1	86.2839846	86.2839846	5.18	0.0267
Error	55	915.8326821	16.6515033		

SAS CODE 2

```
proc glm data=Food; class Treatment;
model Y1_4 Y1_8 Y1_12 Y1_24
      Y2_4 Y2_8 Y2_12 Y2_24
      Y3_4 Y3_8 Y3_12 Y3_24 = Treatment / nouni;
manova h=intercept m= Y1_4  - Y1_8, Y1_8 - Y1_12,
                      Y1_12 - Y1_24, / summary; run;
```

ANALYSING DATA USING THE MIXED PROCEDURE

In repeated measures situations, the mixed model approach, which can be analyzed with PROC MIXED or PROC GLIMMIX, is more flexible and more widely applicable than either the univariate or multivariate approach discussed above. The mixed model approach with PROC MIXED also provides a larger class of covariance structures and a better mechanism for handling missing values (Wolfinger and Chang, 1995) than the traditional univariate or multivariate analysis approach. PROC MIXED is a generalization of the GLM procedure and fits the wider class of mixed linear models, while PROC GLM fits standard linear models. PROC MIXED computes Type I to Type III tests of fixed effects. The CLASS, MODEL, CONTRAST, ESTIMATE, and LSMEANS statements are similar with both procedures, but their REPEATED statement differs. The sorting of classification levels differs between the two although both procedures use the non-full-rank model parameterization. Covariance structures for repeated measurements on subjects are used with the REPEATED statement in PROC MIXED, while the traditional univariate or multivariate tests with various transformations are conducted with the REPEATED statement in PROC GLM. Also, PROC MIXED can perform a sampling-based Bayesian analysis through the PRIOR statement, and supports certain Kronecker-type covariance structures. However, in linear mixed models that are fitted with PROC MIXED, the data

are assumed normally distributed, given the random effects. Whenever, the data contains non-normal rresponse variable a transformation is required. The data used in this study contained response variables from several distributions and not normal. The combined response variable which belonged to lognormal distribution was Box-Cox transformed into normal.

ARRANGING INDEPENDENT AND DEPENDENT VARIABLE IN UNIVARIATE FORMAT

The data with multiple responses need to be arranged into the "tall" univariate format to analyze using PROC MIXED. The variable, Time (4,8,12 and 24 week) and other independent variables were re-arranged to be in the univariate long format (SAS CODE 3). Few observations read from the univariate tall file format created by SAS CODE 3 is shown in Table 4. The response variable was included as a CLASS variable (named VAR in this paper) to fit a multivariate model with the MIXED procedure. The VAR variable in the data which is a combination of three response variables generated three design matrix columns corresponding to three intercept terms, one for each response. The NOINT option was used in the MODEL statement to prevent PROC MIXED from generating another, unnecessary intercept column. In general, VAR is crossed with each other effect in the model to create the required matrix. The SOLUTION or the S option was used with the MODEL statement to output the estimated regression coefficients. An appropriate covariance structure matrix (unstructured in this case) among the several covariance matrices available with PROC MIXED was specified with the REPEATED statement, after being tested for their significance among the several covariance structures. Other aspects of PROC MIXED's input and output apply.

SAS CODE 3

```
/* A variable time was created for the effect of tie */
data times;
input time1-time4;
datalines;
 4 8 12 24
;
Data Tallfood; Set Food;
VAR='Y1'; Y=Y1; output;
VAR='Y2'; Y=Y2; output;
VAR='Y3'; Y=Y3; output;
keep ID Treatment VAR Y;

Data Tallfood1; set Tallfood;
if _n_ =1 then merge times;
array t(4) time1 - time4;
array v(4) VAR_4 VAR_8 VAR_12 VAR_24;
array trt(4) treatment4 treatment8 treatment12 treatment24;
array Y(4)  Y_4 Y_8 Y_12 Y_24;

do i=1 to 4;
week=t(i); var=v(i);
treatment1=trt(i);
response = Y(i);
output; end;
keep ID week Var Response Y Treatment;
RUN;
```

BOX-COX TRANSFORMATION OF DEPENDENT VARIABLE

The response variable was tested for normality (Figure 1). The response variable used in this study was not normal Box Cox transformation was performed with the TRANSREG procedure to transform the response variable into normal. The scatter plot and Box Cox transformation plot of the response variable are shown in Figures 2 and 3 respectively. The Box Cox analysis is shown in Figure 4. See SAS CODE 4 for details of Box-Cox transformation.

SAS CODE 4

```
proc transreg data=tallfood ss2 details
    plots=(transformation(dependent) scatter
               observedbypredicted);
    model BoxCox(y / lambda=-2 -1 -0.5 to 0.5 by 0.05 1 2
    convenient parameter=2 alpha=0.00001) =identity(pid);
     Label Y = 'Response';
  run;
```

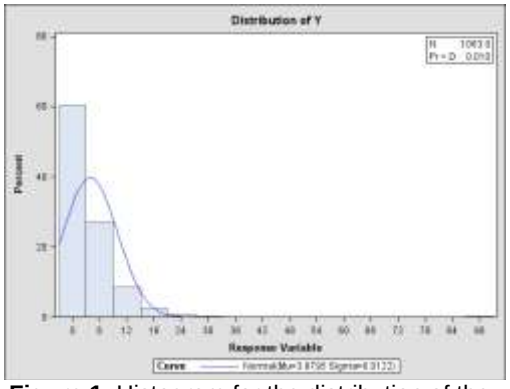

Figure 1: Histogram for the distribution of the overall response variable

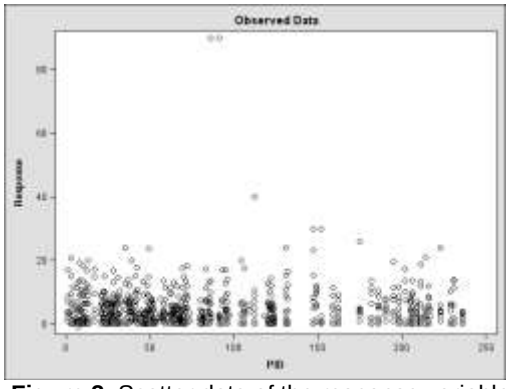

Figure 2: Scatter data of the response variable

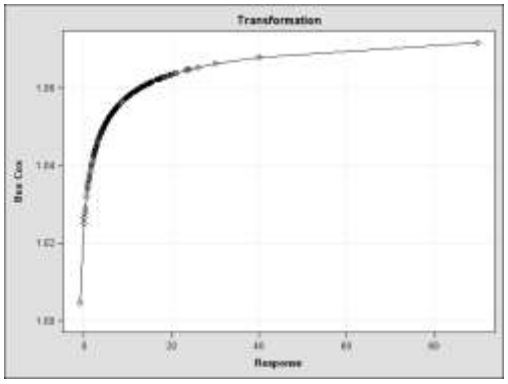

Figure 3: Box-Cox transformed response variable

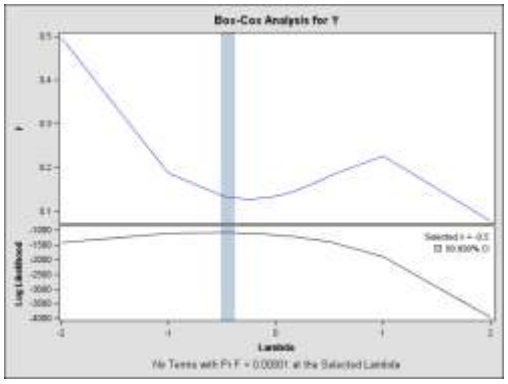

Figure 4: Box-Cox transformation analysis of the response variable

NON ITERATIVE MAXIMUM LIKELIHOOD ANALYSIS

Repeated measures analysis was performed with an unstructured within-subject variance-covariance matrix using PROC MIXED (SAS CODE 5). A residual variance was not profiled in the model. A non-iterative influence analysis was performed to update the fixed effects. The analysis does not take into account the effect on the covariance parameters when a subject is removed from the analysis. If the covariance parameters are updated, the impact of observations on these can amplify or allay their effect on the fixed effects. To assess the overall influence of subjects on the analysis and to compute separate statistics for the fixed effects and covariance parameters, an iterative analysis was performed by adding the sub-option ITER with the INFLUENCE statement. The convergence criteria were satisfied without much effort or taking long computer time.

The analysis incorporates correlations for all of the observations arising from the same person. The transformed data used was Gaussian, and the likelihood was maximized to estimate the model parameters. The null

model likelihood ratio test (Table 6) was highly significant, indicating that the unstructured covariance matrix was preferred to the diagonal matrix of the ordinary least squares null model. The degrees of freedom for this test is 9, which is the difference between 10 and the 1 parameter for the null model's diagonal matrix

Table 4: Twelve observations taken from the middle part of the univariate data created by SAS CODE 3

Obs.	PID	Treatment	Week	Var	Y
1165	223	LC	4	Y1	2.50
1166	223	LC	4	Y2	1.00
1167	223	LC	4	Y3	1.00
1168	223	LC	8	Y1	3.80
1169	223	LC	8	Y2	1.00
1170	223	LC	8	Y3	6.00
1171	223	LC	12	Y1	4.10
1172	223	LC	12	Y2	1.00
1173	223	LC	12	Y3	4.00
1174	223	LC	24	Y1	3.70
1175	223	LC	24	Y2	1.00
1176	223	LC	24	Y3	24.00

SAS CODE 5

```
ods graphics on;
proc mixed data=Tallfood method=ML IC;
class Week VAR Treatment;
model Y = Week VAR Treatment VAR*treatment Week*VAR Week*Treatment
              / NOINT NOTEST S influence(effect=ID ITER=5);
repeated / type=un subject=ID;

ods select influence; run;
```

The REPEATED statement contains no effects, taking advantage of the default assumption that the observations are ordered similarly for each subject. The unstructured covariance matrix was output for each SUBJECT=person (specified by ID). The matrix is, therefore, block diagonal with 27 blocks, each block consisting of identical 44 unstructured matrices. The 10 parameters of these unstructured blocks make up the covariance parameters estimated by maximum likelihood. Solutions to the main fixed effects are shown in Table 7. According to Table 7, time effect is highly significant. The treatment factor, LC is negatively correlated to the response and is significant. The rest of the results are not shown due to space restrictions.

Table 5: Information on variables used with CLASS statement

Class Level Information		
Class	**Levels**	**Values**
Var	3	Y1 Y2 Y3
Week	4	4 8 12 24
Treatment	2	LC LF

Table 6: Likelihood ratio test for unstructured covariance matrix

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
5	138.30	<.0001

Table 7: Solution to fixed effects (shown for the main effects only)

Effect	Var	Week	Treatment	Estimate	Standard Error	DF	t Value	Pr > |t|
Week		**4**		1.3412	0.2295	100	5.84	<.0001
Week		**8**		1.6457	0.2499	100	6.59	<.0001
Week		**12**		3.0971	0.4632	100	6.69	<.0001
Week		**24**		8.9518	1.6446	100	5.44	<.0001
Treatment			**LC**	-0.1518	0.2633	100	-0.58	0.5655
Treatment			**LF**	0	.	.	.	.
Var	**Y1**			1.4114	1.8326	100	0.77	0.4430
Var	**Y2**			-8.0244	1.6410	100	-4.89	<.0001
Var	**Y3**			0	.	.	.	

ANALYSIS WITH THE GLIMMIX PROCEDURE

The GLIMMIX procedure fits generalized linear mixed models (GLMMs). Linear mixed models are in the class of GLMMs. The GLIMMIX procedure accommodates non-normal data and offers a broader array of post-processing features than the MIXED or GLM procedures. Repeated measures analysis was also performed In this study for the data using PROC GLIMMIX and compared to the results obtained from PROC MIXED. If the inference about the treatment over time on the response and their interaction is of the main interest, the changes in the response variables over the 24 week period need to be accounted for in the analysis. A reasonable approach is to apply the approximate low-rank smoother to capture the trends in the response variables over time. This approach avoids the needs to stipulate a parametric model for the response trajectories over time. In addition, hypotheses about the smoothing parameter can be captured; for example, whether it should be varied by treatment. The data were arranged into the long format as described above in section 2.2.1 and was analyzed with PROC GLIMMIX. Since the treatment over time on the response was the major interest in this study, PROC GLIMMIX was found more appropriate for analysis of data used in this study. A parametric model with a factorial treatment structure and smooth trends over time, choosing the Newton-Raphson algorithm with ridging for the optimization was fitted using PROC GLIMMIX (SAS CODE 6). The continuous time effect was included in both the MODEL statement and the RANDOM statement. Since the variance of the radial smoothing component depends on the temporal metric, the time scale needed to be rescaled for the RANDOM effect to move the parameter estimate away from the boundary. The knots of the radial smoother are selected as the vertices of a *k-d* tree, specifying BUCKET=100. The KNOTINFO keyword of the KNOTMETHOD= option provides a printout of the knot locations for the radial smoother. An OUTPUT

statement was used to save the predictions of the mean of each observation to an output data set (gmxout). The TECH=NEWRAP option in the NLOPTIONS statement was used to specify the Newton-Raphson algorithm for the optimization technique to fasten the optimization. A table with the dimensions that contains the G-side variance of the spline coefficients and the R-side scale parameter were output. Output tables with, Optimization Information (not shown), Fit Statistics (not shown), Covariance Parameter Estimates (Table 8) and Type III Tests of Fixed Effects (Table 9) were output too. The fit statistics were found adequate. Plots of smoothed subject specific trends in the response variable over time are shown in Figure 5. The GLIMMIX procedure processes the data by subjects. No variations were observed by subjects. The differences in the effects of the treatment were plotted separately for each of the three response variables in different panels, using PROC SGPANEL for ease of viewing. Predicted unplanned meals (Y3) increased more in the low carbohydrate (LC) group than in the low fat (LF) treatment group. The treatment groups did not differ in their weight loss (Y1) or in their free food intake significantly (Figure 5) over time.

SAS CODE 6

```
proc glimmix data=tallfood; tpoint = week / 100;
class treatment VAR Week;
model Y = Week VAR treatment VAR*treatment VAR*week;
random tpoint / type=rsmooth subject=PID
         knotmethod=kdtree(bucket=100 knotinfo);
output out=gmxout pred(blup)=pred; nloptions tech=newrap; run;
```

Table 8: Covariance parameter estimates from PROC GLIMMIX

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
Var[RSmooth(tpoint)]	PID	677.05	166.72
Residual		21.4676	1.0462

Table 9: Solution to Type III tests of Fixed effects

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Week	3	673	32.71	<.0001
Var	2	673	155.30	<.0001
Treatment	1	673	2.38	0.1234
Treatment*Var	2	673	2.57	0.0770
Var*Week	6	673	15.28	<.0001

SPECIFYING SEPARATE DISTRIBUTION FOR THE THREE RESPONSE VARIABLES

As discussed above the data contained three response variables from three different distributions. The Y2 was a binary response variable (Free Food, if the patient took free food (=1) or not (=0)). The Y1 (weight loss in kilogram) was a continuous response variable with a log normal distribution. The Y3 (number of unplanned meals

consumed) was a count response variable with a Poisson distribution. PROC GLIMMIX allows three different distributions specified for the three response variables which were achieved with a character variable (dist) to identify the distribution of each response variable separately in programming statement. A multivariate logistic analysis was considered for the binary response variable. A multivariate Poisson analysis was used for the count response variable. The count, continuous and binomial response variables were modeled jointly (SAS CODE 7) in PROC GLIMMIX, with the help of the BYOPS option used in the MODEL statement, in PROC GLIMMIX. The variable 'dist' was included to specify the distribution assumed for the three different response variable.

SAS CODE 7

```
proc glimmix data=Tallfood;
class ID dist Var Week;
model response(event='1') = dist dist*tretment dist*Y /
              noint s dist=byobs(dist);
random int / subject=ID;
```

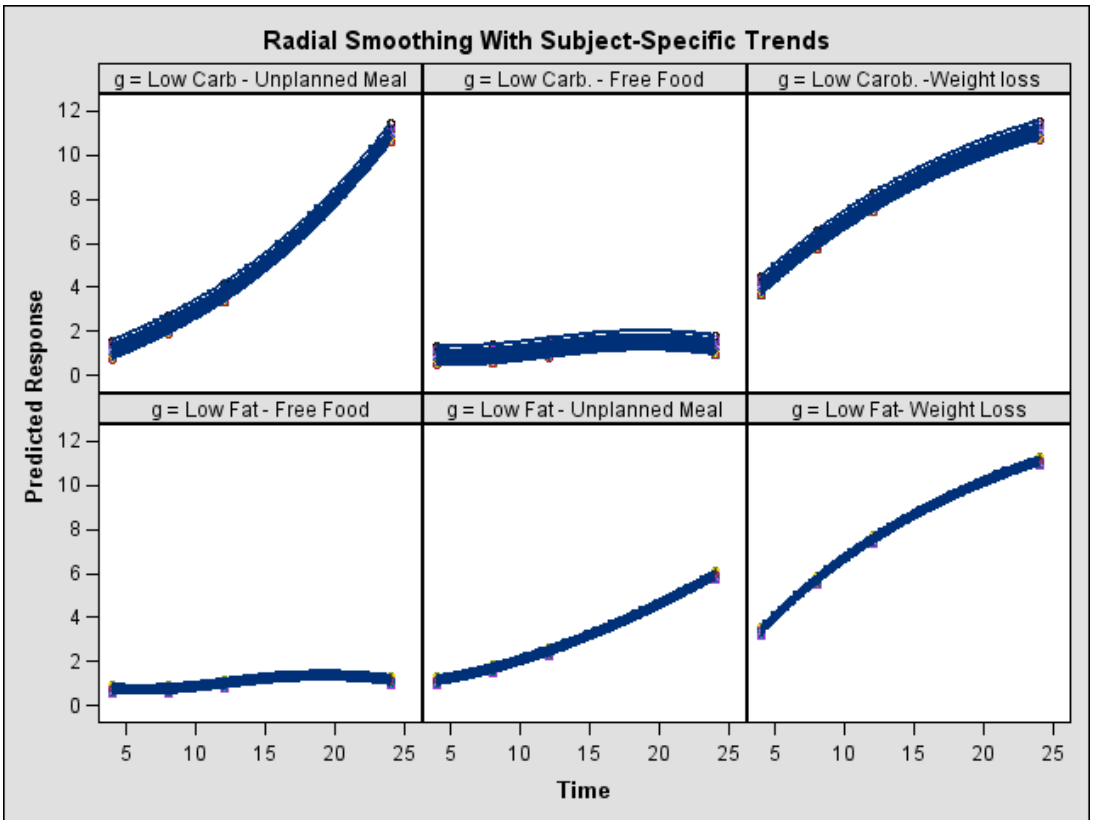

Figure 5: Subject specific trends of the predicted response

INCORPORATING CORRELATION AMONG RESPONSE VARIABLES

The correlation among the three response variables for the same patient was incorporated in the model through modeling the dependency directly into the model (SAS CODE 8). The ID variable was added to the CLASS statement and as the SUBJECT= effect in the RANDOM statement to further take care of the correlation among the response variables.

SAS CODE 8

```
Data Tallfood; Set Food; length dist $7;
response= (Y2=2); dist = "Binary"; output;
response= (Y2=3); dist = "Poisson"; Output;
response= (Y2=1); dist = "LogNormal"; Output;
keep ID week treatment response dist; Run;

class dist Var Treatment Week;
model response(event='1') = dist Var Treatment Week
treatment*dist / s dist=byobs(dist);
```

CONCLUSIONS

The correlation among repeated measurements in subjects can be taken into account by performing a repeated measures analysis of variance. In addition, when there are multiple responses, a multivariate analysis combined with tools of repeated measures is required, which can only be achieved by careful manipulation in programming structure. In addition, careful attention to several details such as distributional variation among responses, correlations among responses, subject specific trend in the treatment variable need to be taken care of which is a challenging task for a SAS programmer. In this research, repeated measures analysis of correlated data with multiple response variables that are a mixture of continuous, count, and binomial was successfully performed using SAS. The common problems that arise when analyzing such data were addressed in detail in SAS. Three SAS procedures were compared. PROC GLIMMIX was found to be superior to PROC MIXED since it allowed inclusion of distributional variation among response variable, and radial smoothing of subject specific trend in treatment over time.

REFERENCES

Wolfinger, R.D. and Chang, M. (1995). Comparing the SAS GLM and MIXED procedures for repeated measures. *Proceedings of the Twentieth Annual SAS Users Group Conference*, SAS Institute Inc., Cary, NC.

CONTACT THE AUTHORS

Most details of the analysis and SAS output were omitted from this paper due to space restriction. Please contact the author if any details on any of the analysis are required.

Name:	Anpalaki J Ragavan	Anbuchelvi Jeyabalasingham
Address:	1205 Beech Street, Apt. 11, Reno, NV 89512, USA	621, Jeyabal Mansion, Velikkulam, Vavuniya, Sri Lanka
Phone -Home:	775-322-3694	011-94-24-2220406
Pnone-Cell	775-327-5260	011-94-774125193
Email:	ragavan@unr.edu	anpalakir@unr.edu
Website:	None	None