

Using SAS/STAT® To Analyze Correlated Data

Marlene Goormastic
Cleveland Clinic Foundation
Cleveland, Ohio

Correlated data can result from several different study designs. Conceptually we can think of correlated data as falling into one of two categories; repeated measures and clustered data. Repeated measurements involve the same factor in the same experimental unit being measured more than once. These measurements may occur at the same time or be taken over time. There may or may not be an intervening event which is expected to impact the measurement. Clustered data results from two situations. The first is when measurements on similar but not identical factors are measured on the same experimental unit (e.g. vision acuity in both eyes). The second is when the same factor is measured on different but related experimental units (e.g. birth weight in a litter of pup or a survey of a family household).

Since correlated data violates one of the most fundamental assumptions underlying most statistical tests, that of independence of the observations, its analysis has always been more problematical. Even the common situation of two repeated measurements, e.g. pre and post procedure, has to be handled indirectly in SAS by creating the difference and running a PROC UNIVARIATE and requesting the t-test or signed rank test. Recent enhancements in existing PROCs and the addition of PROC MIXED in version 6.07 has provided many more options for dealing with correlated data. Examples of data analyzed using each of the following; PROC GLM, PROC GENMOD and PROC MIXED are presented in this paper.

GLM

This procedure fits least square means for general linear models.

PROC GLM can be used in a variety of situations with repeated data. Such as the previous pre-post situation (paired t-test) if there are other factors to adjust for, more than two groups, or more than two repeated measurements.

Example 1A:

Thirty-six rabbits underwent coronary angioplasty and had the extent of disease measured pre, post and at 4-6 weeks.

In this case there is no grouping factor but this can be handled by creating a dummy grouping variable which has the same value for all the cases. This enables you to test for the time effect.

```
data new;
input obs pre post week3;
group=1;
```

cards;

```
1 1.12 1.64 0
2 1.37 1.74 0.78
3 1.38 1.73 1.11
4 1.31 1.76 1.54
5 0.89 1.59 0.76
6 1.31 1.64 1.24
7 1.42 1.78 0.31
8 1.23 1.78 0.66
etc.
```

```
PROC GLM data=new; class group;
model pre post week1=group/nouni;
repeated time 3/printe;
```

```
run;
```

The nouni option suppresses the univariate results for each of the dependent variables (pre post and week3) which are usually of no interest in a repeated measures analysis. Printe prints the E matrix for each combination of within-subject factors.

The three time measurements are listed in the model statement. The repeated statement includes a name of your choice to describe how the measurements are repeated. "Time" was used in this program. This is followed by the number of repeated measurements. Since GLM requires that all the repeated measurements be in the same record, the subject is not specified. The format of the data is what is referred to as multivariate format, multiple dependent variables are listed in the model. Excerpts from the SAS output for these examples are provided at the end of this paper in Appendix 1.

The program would look identical if there had been a real grouping factor.

GLM can also handle more than one repeated factor.

Example 1B:

Measurements are taken at 2 gravity levels (2 and 5) and 2 times (2 and 10 minutes).

```
options errors=1 nocenter ls=80 ps=66;
proc format; value grp 1='Donor' 2='Patient';
```

```
data three;
input grp id $ baseline g2m2 g2m10 g5m2 g5m10;
cards;
1 d007 0.607 0 0 0.088 0
```

```
1 d002 0.0024 1.049 0.083 0 0
1 d150 1.566 1.494 3.232 7.898 7.766
1 d001 21.62 71.827 228.007 68.097 963.107
1 d003 1.376 8.639 0.075 0 0.307
1 d005 1.272 0.507 20.283 3.576 38.233
1 d009 0 2.126 3.266 6.006 3.226
2 td 0.6 0 0 0 .
2 ld 11.233 65.4 184.133 . .
2 re 0.605 0 3.794 0 0.211
2 bg 27.047 49.047 116.617 151.717 174.217
2 sb 0 1.532 2.128 1.069 3.439
2 mp 0.477 5.556 14.966 16.726 16.736
2 rc 0 7.525 3.221 6.005 4.764
2 ms 0.365 0 0.142 0.851 0.814
etc.
```

```
*****
PROC GLM data=three; class grp;
  model g2m2 g2m10 g5m2 g5m10=grp /nouni;
  repeated grav 2 (2 5) time 2 (2 10) /short;
run;
*****
```

The output (see Appendix 1) tells us that SAS correctly recognizes the factors being repeated. There was also a baseline value which you may want to adjust for. One way is to subtract it from each of the four measurements and run the above model on the adjusted variables. Another approach is to include it in a separate analysis including the baseline for gravity and time.

```
*****
PROC GLM data=three; class grp;
  model baseline g2m2 g2m10=grp /nouni;
  repeated time 3 (0 2 10) /short;
run;
```

This would test the time effect, including baseline at a fixed gravity level (2).

Often the correlated data is in a "univariate" format with a separate record for each measurement, such as the following:

visit	id	response
0	1	1.12
1	1	1.64
3	1	1.70
0	2	1.37
1	2	1.73

etc.

This format is more efficient since no record is entered for missing observations. Plotting the response over time is also straightforward. Unfortunately, data in this format has to be transformed before GLM can be used (see Proc Transpose).

GENMOD

This procedure is an extension of generalized linear models where the mean of the data depends on a linear predictor through a link function. This enables the response probability distribution to be any member of the exponential family of distributions. In other words, PROC GENMOD can be used to analyze correlated data where the outcome is binary. Correlated data is handled by using Generalized Estimating Equations (GEEs). GEEs can be fit using the repeated option.

Example 2:

Heart Transplant patients have an diagnostic test to measure disease in the coronary arteries within 1 month post transplant. A year later the test is repeated but in addition to the multiple segments measured at baseline, additional segments are measured. The questions are, 1) does the diagnostic test itself injure the arteries and promote disease? and 2) Is the disease in the segments that have been imaged twice worse than take of the segments imaged once? Other factors to consider are the coronary artery type of the segment (RCA, LAD, CX) and the location of the segment within that artery (proximal, mid or distal). Disease can be analyzed as the actual amount of disease or classified as present or not (0,1).

In addition you can specify the distribution (e.g. normal or binomial) and the covariance matrix structure:

- independent (no correlation),
- auto regressive (correlation lessens as observations become more distance from each other)
- exchangeable (the correlation between any two observations is the same)
- and unstructured (estimated from the data).

Or the user can supply their own correlation matrix.

```
options errors=1 nocenter ls=80 ps=66;
```

```
data xxx; input id ves $ seg_id $ paired amount;
  if seg_id in ('19' '19A' '19B' '19C') then site=3; *distal;
  else if seg_id in ('18' '18A' '18B' '18C') then
    site=1; *prox;
  else if seg_id in ('20' '20A' '20B' '20C') then
    site=1; *mid;
  if amount>=0.5 then dx=1; else dx=0;
  if seg_id in ('1' '1A' '1B' '2' '2A' '2B' '3' '3A' '3B'
    '4' '5' '6' '7' '8' '9') then ves_type=1; *RCA;
```

.....more recoding

```
cards;
1065 CX 18A 1 0
1065 CX 18B 1 0.14
1065 RCA 1A 0 0.23
1065 RCA 1B 0 0.29
1065 CX 20A 0 0.17
1065 CX 20B 0 0.08
1918 LAD 12A 1 0.74
1918 LAD 12B 1 0.86
1918 LAD 13A 1 0.31
1918 LAD 13B 0 1.3
1993 RCA 1 0 0.44
```

1993 LAD 12 1 1.61
 1993 LAD 13A 1 0.12
 1993 LAD 13B 1 1.39
 2003 LAD 12A 0 0.51
 2003 LAD 12B 0 1.04
 2003 CX 20A 1 0.17
 2003 CX 20B 1 0.44
 etc.

The program for disease (no, yes).....

```
PROC GENMOD data=xxx;
  class id ves_type ves_loc;
  model dx = paired ves_type ves_loc/dist= bin;
  repeated subject=id /corrw type=exch;
```

run;

dist=bin uses the binomial distribution
 corr=exch uses an exchangeable correlation structure
 corrw prints the estimated working correlation matrix

this program is using the amount of disease as the dependent variable

```
PROC GENMOD data=xxx;
  class id ves_type ves_loc;
  model amount = paired ves_type ves_loc/dist= nor;
  repeated subject=id /corrw type=exch;
```

run;

dist=nor uses a normal distribution

MIXED

This procedure fits models with both fixed and random effects (mixed linear models). It is a generalization of the standard linear model used in GLM in that more than one source of variation can be handled. Simplistically speaking, random effects are factors that you need to control for but aren't really interested in while the fixed effects are those factors whose impact you'd like to test for. In the repeated measures situation, the subject is usually considered the random effect that you want to control or adjust for.

Example 3:

Ten morbidly obese patients underwent gastric bypass surgery. Weight measurements were taken at several follow-up visits. The first five patients had a suture technique while the remaining five had staples. Was there significant weight loss over time and did it differ by group? As the plot of the data shows, it is the slopes that are of interest. To model the slopes in this situation the random statement rather than the repeated statement is used. The random statement will fit slopes with different intercepts.

data fup; input id 2. visitnum 2. @6 wt;

cards;
 1 0 148
 1 1 137
 1 2 136
 1 3 132.5
 1 4 125
 1 5 119
 1 6 114
 1 7 105
 2 0 173
 2 1 134.5
 2 2 128
 2 4 111
 2 5 99
 3 0 152
 3 2 135
 4 0 142
 4 2 122.5
 4 3 118
 4 4 112
 4 5 110
 4 6 102
 etc.

Slope Model

```
PROC MIXED data=fup;
  class id grp;
  model weight = visitnum grp visit*grp /s;
  random int visitnum/ subject=id type=un;
end;
```

type=un requests an unstructured covariance block for each person

ANOVA Model

```
PROC MIXED data=xxx;
  class id grp;
  model weight = visitnum grp visit*grp /s;
  repeated visitnum/ subject=id type=cs;
  *****
```

The unstructured covariance would not converge under the ANOVA model. For this example a compound symmetric structure (a common covariance with a diagonal enhancement) was specified instead.

Author Information:

Marlene Goormastic, MPH
 Lead Biostatistician, Transplant Center
 Cleveland Clinic Foundation
 Cleveland, Oh 44195
 mgoormas@bio.riccf.org

References:

GLM

SAS/STAT® User's Guide, Version 6, Fourth Edition Volume 2 p951-996.

MIXED

SAS® Technical Report P-229, SAS/STAT® Software: Changes and Enhancements, Release 6.07
Chapter 16 The Mixed Procedure p287-364

GENMOD/GEE

Stokes, Maura. Statistical Enhancements in Release 6.12 of the SAS® System

Johnston Gordon and Stokes, Maura. Repeated Measures Analysis with Discrete Data Using the SAS® System
Proceedings of the Twenty-Second Annual SAS® Users Group International Conference,
p1293-1299 and p1300-1305.

APPENDIX 1

-----excerpt from SAS OUTPUT Example 1A-----

Rabbit Study

General Linear Models Procedure
Class Level Information

Class	Levels	Values
GROUP	1	1

Number of observations in data set = 36

Dependent Variable	PRE	POST	WEEK3
Level of TIME	1	2	3

General Linear Models Procedure
Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

Source: TIME						Adj	Pr > F
DF	Type III SS	Mean Square	F Value	Pr > F	G - G	H - F	
2	16.30231852	8.15115926	57.43	0.0001	0.0001	0.0001	

Source: TIME*GROUP						Adj	Pr > F
DF	Type III SS	Mean Square	F Value	Pr > F	G - G	H - F	
0	.	.	.	.	.	.	

Source: Error (TIME)		
DF	Type III SS	Mean Square
70	9.93548148	0.14193545

Greenhouse-Geisser Epsilon = 0.8235
Huynh-Feldt Epsilon = 0.8589

----- excerpt from SAS OUTPUT Example 2 -----

General Linear Models Procedure
Class Level Information

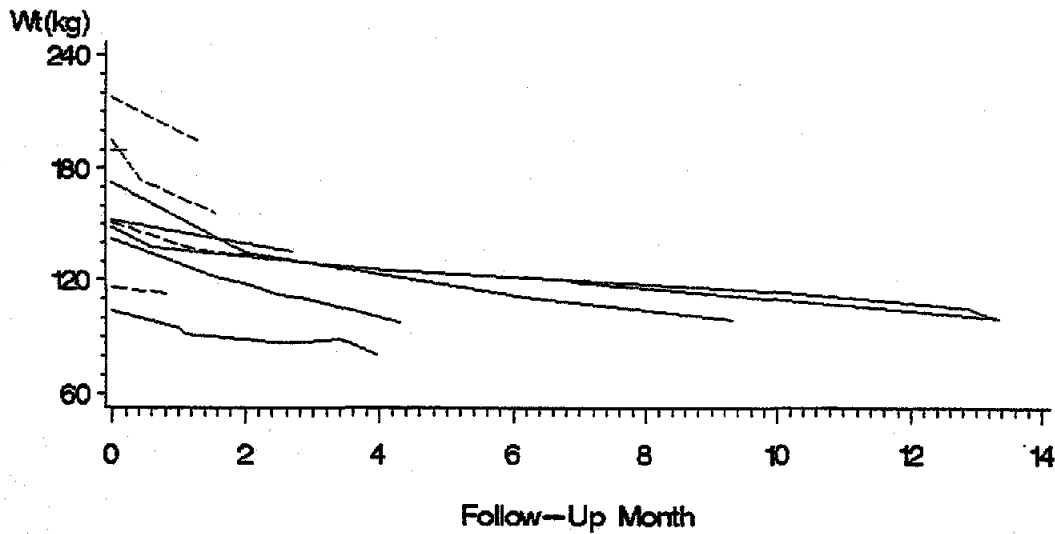
Class	Levels	Values
GRP	2	Donor Patient

General Linear Models Procedure
Repeated Measures Analysis of Variance
Repeated Measures Level Information

Dependent Variable	G2M2	G2M10	G5M2	G5M10
Level of GRAV	2	2	5	5
Level of TIME	2	10	2	10

Plot Of Data For Example 3

Morbid Obesity Study
Change in Wt(kg) Over Time



solid lines = suture, dashed lines = staples

-----Excerpt from SAS OUTPUT Example 3 Slope Model-----

```

The MIXED Procedure
Class Level Information
Class      Levels  Values
GRP        2      1st 2nd
          G Matrix
Effect      Subject  Row      COL1      COL2
INTERCEPT      1      1 1080.0476579 -100.1818716
VISITNUM         1      2 -100.1818716  16.00542469

```

Covariance Parameter Estimates (REML)

```

Cov Parm  Subject      Estimate
UN(1,1)    NID      1080.0476579
UN(2,1)    NID     -100.1818716
UN(2,2)    NID      16.00542469
Residual                25.47398560

```

Solution for Fixed Effects

Effect	GRP	Estimate	Std Error	DF	t	Pr > t
INTERCEPT		174.10044374	14.85163595	8	11.72	0.0001
VISITNUM		-10.55601001	2.66615242	8	-3.96	0.0042
GRP	1st	-34.87215147	20.96275120	22	-1.66	0.1104
GRP	2nd	0.00000000	.	.	.	.
VISITNUM*GRP	1st	3.34582609	3.27528722	22	1.02	0.3181
VISITNUM*GRP	2nd	0.00000000	.	.	.	.

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
VISITNUM	1	8	29.42	0.0006
GRP	1	22	2.77	0.1104
VISITNUM*GRP	1	22	1.04	0.3181

-----excerpt from SAS OUTPUT Example 3 ANOVA model-----

```

The MIXED Procedure
Class Level Information
Class      Levels  Values
GRP        2      1st 2nd
VISITNUM   0 1 2 3 4 5 6 7

```

Covariance Parameter Estimates (REML)

```

Cov Parm  Subject      Estimate
CS         NID      871.30173871
Residual                61.50677877

```

R Matrix for Subject 1

Row	COL1	COL2	COL3	COL4	etc.
1	932.80851748	871.30173871	871.30173871	871.30173871	
2	871.30173871	932.80851748	871.30173871	871.30173871	
3	871.30173871	871.30173871	932.80851748	871.30173871	
4	871.30173871	871.30173871	871.30173871	932.80851748	
5	etc.				

Solution for Fixed Effects

Effect	GRP	VISITNUM	Estimate	Std Error	DF	t	Pr > t
INTERCEPT			116.51638035	16.76453811	8	6.95	0.0001
VISITNUM	0		57.48361965	10.91287066	21	5.27	0.0001
VISITNUM	1		47.41040527	11.26460413	21	4.21	0.0004
VISITNUM	2		34.05455960	9.04624690	21	3.76	0.0011
VISITNUM	3		29.69326326	9.36065609	21	3.17	0.0046
VISITNUM	4		23.19177120	9.13371094	21	2.54	0.0191
VISITNUM	5		16.19177120	9.13371094	21	1.77	0.0908
VISITNUM	6		13.28733647	9.82024471	21	1.35	0.1904
VISITNUM	7		6.78733647	9.82024471	21	0.69	0.4970
VISITNUM	8		2.00000000	11.09114771	21	0.18	0.8586
VISITNUM	9		0.00000000				
GRP	1st		-27.97093995	19.64120309	8	-1.42	0.1922
GRP	2nd		0.00000000				
GRP*VISITNUM	1st	0	-2.22906005	7.86504130	21	-0.28	0.7796
GRP*VISITNUM	1st	1	-10.17960132	8.95476278	21	-1.14	0.2684

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
VISITNUM	9	21	12.99	0.0001
GRP	1	8	2.87	0.1288
GRP*VISITNUM	2	21	0.73	0.4936