Analysis of Longitudinal Data: Comparison Between PROC GLM and PROC MIXED

Maribeth Johnson

Medical College of Georgia, Augusta, GA

Longitudinal Data

- Multiple measurements of a response variable
- On the same experimental unit
- Made over a period of time

Correlated Data

- Quantify the inter-relationships of the measurements
- If variances and covariances are constant then the relationship can be thought of as ICC reliability
- Assumption of compound symmetry can be tested against other variance structures using PROC MIXED

Correlated Data

- The ability to model different covariance structures provide an opportunity to investigate and possibly quantify the tracking of measurements as well as provide the basis for other relationships

Data

- Texas site in the studies of child activity and nutrition (SCAN) program
- Multi-center longitudinal study of the role of nutrition and physical activity on the development of cardio-vascular disease risk factors and associated behaviors in families of young children
- Children aged 3 or 4 years

Data Structure

	Number of children	
Number of records	WT	HR
1 of 4	90	114
2 of 4	34	43
3 of 4	36	58
4 of 4	98	60
	257	275

Data Structure

	WT, kg (n=257)			HR, bpm (n=275)		
Year	N	Mean	SD	N	Mean	SD
1986	233	16.1	2.4	249	93.7	10.2
1987	145	18.5	3.0	144	90.4	10.0
1988	143	21.3	4.0	111	87.5	8.3
1989	135	24.2	4.7	110	82.2	9.7

Using GLM With The Random Statement

```
proc glm; class child year;
      model depvar = child year;
      random child;
```

- Mixed model univariate analysis of variance
- Random statement prints table of expected MS

Using GLM With The Random Statement

The GLM Procedure

Source	Type III Expected Mean Square
CHILD	Var(Error) + 4 Var(CHILD)
YEAR	Var(Error) + Q(YEAR)

Using GLM With The Random Statement

- E(MS) Child = Var(Error) + *k* Var(CHILD)

  k is the average number of observations per child

  CHILD is treated as a fixed effect

- $$\text{Var(CHILD)} = \frac{\text{MS(CHILD) - MS(Error)}}{k}$$

- $$\text{ICC} = \frac{\text{Var(CHILD)}}{\text{Var(CHILD) + MS(Error)}}$$

Using GLM With The Random Statement

- Reliability of a single year of measurement
- The individual components of variation are maximum-likelihood estimators when the data are balanced

Using GLM With The Repeated Statement

- Can’t handle missing data
- Repeated measurements must appear in a multivariate mode in the dataset
- Allows for the multivariate test of the assumption of compound symmetry (CS)
- If the CS assumption is rejected it can’t help in determining the correct underlying covariance structure

Using GLM With The Repeated Statement

```
proc glm;
  model depvar1-depvar4 = /nouni;
  repeated year / printe;
```

- NOUNI option suppresses the univariate analyses of each year
- PRINTE option outputs the partial correlations computed from residuals after fitting the between-subjects model

“Test for Sphericity”

- Tests whether a set of orthonormal contrasts of the repeated measures variables are independent and have equal variances, i.e. are the data compound symmetric
- Significance tells you that this condition is not met
- The estimate of the correlation between measures from the univariate GLM analysis is not valid

“Test for Sphericity”

The GLM Procedure

Repeated Measures Analysis of Variance

Sphericity Tests

Variables	DF	Mauchly's Criterion	Chi-Square	Pr > ChiSq
Transformed Variates	5	0.0086796	454.37268	<.0001
Orthogonal Components	5	0.0675075	258.0209	<.0001

“Test for Sphericity”

- Correlations from the PRINTE option of this analysis are **identical** to the correlations computed from the RCORR option in the REPEATED statement in PROC MIXED when TYPE=UN is specified as the covariance structure
- Tests of fixed effects from both analyses are similar
- REPEATED statements perform different functions

Using PROC MIXED

```
proc mixed; class child year;
    model depvar = year;
    repeated year / subject=child r  rcorr
        type=cov-structure ;
```

- REPEATED statement models the covariance structures in **R**, the variance-covariance matrix of the vector of errors

Using PROC MIXED

```
proc mixed; class child year;
     model depvar = year;
     repeated year / subject=child r  rcorr
          type=cov-structure ;
```

- SUBJECT=CHILD option is the mechanism for block diagonalizing **R**

Using PROC MIXED

```
proc mixed; class child year;
    model depvar = year;
    repeated year / subject=child r  rcorr
         type=cov-structure ;
```

- R option of the REPEATED statement requests that the first block of the **R** matrix be printed

Using PROC MIXED

```
proc mixed; class child year;
    model depvar = year;
    repeated year / subject=child r  rcorr
        type=cov-structure ;
```

- RCORR options prints the correlation matrix corresponding to **R**

Using PROC MIXED

```
proc mixed; class child year;
     model depvar = year;
     repeated year / subject=child r  rcorr
          type=cov-structure ;
```

- TYPE= option is what determines the V-C structure

Covariance Structures

- Compound symmetric (CS):
  - Most specific structure
  - Variance within years is constant
  - Common correlation between years
  - Two parameters estimated
  - Assumption of ANOVA estimates

Covariance Structures

- Heterogeneous compound symmetric (CSH):
  - Common correlation
  - Different variances along the diagonal
  - Five parameters estimated

Covariance Structures

- First-order autoregressive (AR(1)):
  - Variance within years is constant
  - Estimates the autoregressive parameter
    - Correlations between years separated by the same amount of time are the same (ρ^m)
  - Two parameters estimated

Covariance Structures

- Heterogeneous first-order autoregressive (ARH(1)):
  - Different variances along the diagonal
  - Estimate of the autoregressive parameter
    - Correlations between years separated by the same amount of time are the same (ρ^m)
  - Five parameters estimated

Covariance Structures

- Unstructured (UN):
  - Estimates of all four variances and six covariances
  - All of the correlations between years may be different
  - Identical to those from the PRINTE option of the REPEATED statement in GLM

Determine The Preferred Model

- **Likelihood Ratio Test (LRT)**
  - One model is a submodel of another
  - Compute -2 times the difference between their residual log likelihoods (-2RLL)
  - Chi-square distribution with degrees of freedom equal to the difference in the number of parameters for the two models
  - Models are preferred where the -2RLL is smaller

Determine The Preferred Model

Akaike's Information Criterion (AIC)

Schwarz's Bayesian Criterion (BIC)

- Model that has the smallest value is the preferred model
- BIC penalizes models with more covariance parameters more than AIC

Determine The Preferred Model

Fit Statistics	
-2 Res Log Likelihood	1657.6
AIC (smaller is better)	1661.6
AICC (smaller is better)	1661.6
BIC (smaller is better)	1666.7

Example
WT Balanced Data

	WT, kg		
Year	N	Mean	SD
1986	98	16.4	2.1
1987	98	18.3	2.6
1988	98	21.2	3.3
1989	98	24.5	4.6

V-C and Correlations WT - Balanced Data (N=98)

Type = CS:

1	2	3	4
10.8	8.9	8.9	8.9
1.0	.82	.82	.82

(Variance and covariances in top line, correlations below)

Type = CSH:

1	2	3	4
4.7	5.0	6.6	9.2
.92	6.4	8.0	10.7
.92	.92	11.0	14.1
.92	.92	.92	21.3

(Variances on diagonal, covariances above, correlations below)

- parameters estimated from the CS structure are identical to those calculated using the univariate GLM mean square estimates

V-C and Correlations WT - Balanced Data (N=98)

Type = CS:

	1	2	3	4
	10.8	8.9	8.9	8.9
	1.0	.82	.82	.82

(Variance and covariances in top line, correlations below)

Type = CSH:

1	2	3	4
4.7	5.0	6.6	9.2
.92	6.4	8.0	10.7
.92	.92	11.0	14.1
.92	.92	.92	21.3

(Variances on diagonal, covariances above, correlations below)

- The sphericity test from the MV GLM was significant

V-C and Correlations WT - Balanced Data (N=98)

Type = CS:

1	2	3	4
10.8	8.9	8.9	8.9
1.0	.82	.82	.82

(Variance and covariances in top line, correlations below)

Type = CSH:

1	2	3	4
4.7	5.0	6.6	9.2
.92	6.4	8.0	10.7
.92	.92	11.0	14.1
.92	.92	.92	21.3

(Variances on diagonal, covariances above, correlations below)

- The correlation from the CSH structure is identical to that calculated using the univariate GLM analysis of WT standardized within year

Model Comparisons WT - Balanced Data

Type	Parameters	-2RLL	Comparison Model	Chi-square / df
CS	2	1658	-	---
CSH	5	1370	CS	288 / 3 *

*p<.0005

- Heterogeneous variances along the diagonal provide a significantly better fit for the CS model

V-C and Correlations WT - Balanced Data (N=98)

Type = AR(1):

1	2	3	4
12.2	11.3	10.5	9.7
1.0	.93	.86	.80

(Variance and covariances in top line, correlations below)

Type = ARH(1):

1	2	3	4
4.5	5.1	6.4	8.4
.95	6.5	8.1	10.6
.91	.95	11.3	14.8
.86	.91	.95	21.4

(Variances on diagonal, covariances above, correlations below)

- The correlations do not fall off as quickly when variances are allowed to differ

Model Comparisons WT - Balanced Data

Type	Parameters	-2RLL	Comparison Model	Chi-square / df
AR(1)	2	1518	--	---
ARH(1)	5	1292	AR(1)	226 / 3 *

*p<.0005

- Heterogeneous autoregressive covariance structure (ARH(1)) provides a better fit than the assumption that within year variances are equal (AR(1))

V-C and Correlations WT - Balanced Data (N=98)

Type = CSH:

1	2	3	4
4.7	5.0	6.6	9.2
.92	6.4	7.7	10.7
.92	.92	11.0	14.1
.92	.92	.92	21.3

Type = ARH(1):

1	2	3	4
4.5	5.1	6.4	8.4
.95	6.5	8.1	10.6
.91	.95	11.3	14.8
.86	.91	.95	21.4

Type = UN:

1	2	3	4
4.6	5.3	6.5	8.4
.96	6.6	8.1	10.6
.90	.95	11.1	14.3
.86	.91	.94	20.8

- UN produced very similar results to ARH(1)
- UN identical to the correlations from the multivariate GLM analysis

Model Comparisons WT - Balanced Data

Type	Parameters	-2RLL	Comparison Model	Chi-square / df
CSH	5	1370	--	---
ARH(1)	5	1292	--	---
UN	10	1284	CSH	86 / 5 *
UN	10	1284	ARH(1)	8 / 5

*p<.0005

- UN provides a significantly better fit to the data than CSH
- ARH(1) appears to provide the best fit for these data
- Improper models may underestimate the correlation between adjacent measurements

Example
WT Unbalanced

		WT, kg	
Year	N	Mean	SD
1986	233	16.1	2.4
1987	145	18.5	3.0
1988	143	21.3	4.0
1989	135	24.2	4.7

V-C and Correlations WT - Unbalanced Data (N=257)

Type = CS:

1	2	3	4
11.5	9.5	9.5	9.5
1.0	.83	.83	.83

(Variance and covariances in top line, correlations below)

Type = CSH:

1	2	3	4
6.1	6.8	9.4	12.2
.94	8.6	11.1	14.4
.94	.94	16.4	19.8
.94	.94	.94	27.5

(Variances on diagonal, covariances above, correlations below)

- ANOVA estimates from GLM no longer ML estimators, not identical to CS but similar
- Var: 12.2 vs 11.5 Cov: 10.1 vs 9.5

Model Comparisons WT - Unbalanced Data

Type	Parameters	-2RLL	Comparison Model	Chi-square / df
CS	2	3012	--	---
CSH	5	2601	CS	411 / 3 *

*p<.0005

- Heterogeneous variances along the diagonal provide a significantly better fit for the CS model

V-C and Correlations WT - Unbalanced Data (N=257)

Type = AR(1):

	1	2	3	4
	12.1	11.1	10.1	9.3
	1.0	.91	.84	.77

(Variance and covariances in top line, correlations below)

Type = ARH(1):

1	2	3	4
6.0	7.0	9.2	11.8
.96	8.9	11.8	15.0
.93	.96	16.8	21.4
.89	.93	.96	29.5

(Variances on diagonal, covariances above, correlations below)

- The correlations do not fall off as quickly when variances are allowed to differ

Model Comparisons WT - Unbalanced Data

Type	Parameters	-2RLL	Comparison Model	Chi-square / df
AR(1)	2	2789	--	---
ARH(1)	5	2505	AR(1)	284 / 3 *

*p<.0005

- Heterogeneous autoregressive covariance structure (ARH(1)) provides a better fit than the assumption that within year variances are equal (AR(1))

V-C and Correlations
WT - Unbalanced Data (N=257)

Type = CSH:

1	2	3	4
6.1	6.8	9.4	12.2
.94	8.6	11.1	14.4
.94	.94	16.4	19.8
.94	.94	.94	27.5

Type = ARH(1):

1	2	3	4
6.0	7.0	9.2	11.8
.96	8.9	11.8	15.0
.93	.96	16.8	21.4
.89	.93	.96	29.5

Type = UN:

1	2	3	4
6.0	7.1	9.0	11.5
.97	8.9	11.6	15.0
.90	.96	16.5	21.2
.87	.93	.96	29.3

- UN produced very similar results to ARH(1)

Model Comparisons WT - Unbalanced Data

Type	Parameters	-2RLL	Comparison Model	Chi-square / df
CSH	5	2601	--	---
ARH(1)	5	2505	--	---
UN	10	2496	CSH	105 / 5 *
UN	10	2496	ARH(1)	9 / 5

*p<.0005

- UN provides a significantly better fit to the data than CSH
- ARH(1) appears to provide the best fit for these data

V-C and Correlations WT – Preferred Models

Type = ARH(1): Balanced data

1	2	3	4
4.5	5.1	6.4	8.4
.95	6.5	8.1	10.6
.91	.95	11.3	14.8
.86	.91	.95	21.4

Type = ARH(1): Unbalanced data

1	2	3	4
6.0	7.0	9.2	11.8
.96	8.9	11.8	15.0
.93	.96	16.8	21.4
.89	.93	.96	29.5

- Correlation estimates are similar
- Variance and covariance estimates higher for unbalanced data
- Anywhere from 37 to 145 more children measured in any one year

Example
HR Balanced Data

		HR, bpm	
Year	N	Mean	SD
1986	60	94.3	8.9
1987	60	91.1	9.3
1988	60	86.8	7.9
1989	60	81.0	9.9

V-C and Correlations HR – Balanced Data (N=60)

- ANOVA estimates and those using the CS structure are identical (r=0.55)
- Common correlation estimate is similar using CSH (r=0.56)
- Heterogenous variances do not provide a better fit for these data
- The MV GLM's test of sphericity was significant so CS is not the correct structure

V-C and Correlations HR - Balanced Data (N=60)

Type = UN:

1	2	3	4
78.6	49.1	25.2	41.9
.60	85.6	42.0	59.2
.36	.58	62.1	51.9
.48	.65	.67	98.1

(Variances on diagonal, covariances above, correlations below)

- The correlation between successive years is lower when the children are young

Model Comparisons WT - Balanced Data

Type	Parameters	-2RLL	Comparison Model	Chi-square / df
CS	2	1639	--	---
AR(1)	2	1643	--	---
UN	10	1622	CS	17 / 8 *
UN	10	1622	AR(1)	21 / 8 *

*p<.05

- UN provides a significantly better fit to the data than CS
- UN provides a significantly better fit to the data than AR(1)

Example
HR Unbalanced

	HR, bpm (n=275)		
Year	N	Mean	SD
1986	249	93.7	10.2
1987	144	90.4	10.0
1988	111	87.5	8.3
1989	110	82.2	9.7

V-C and Correlations HR – Preferred Models

Type = UN: Balanced data

1	2	3	4
78.6	49.1	25.2	41.9
.60	85.6	42.0	59.2
.36	.58	62.1	51.9
.48	.65	.67	98.1

Type = UN: Unbalanced data

1	2	3	4
105.7	56.8	30.4	44.0
.56	98.0	43.7	54.5
.36	.54	68.0	53.2
.45	.58	.68	90.1

- Parameter estimates are similar
- Variance smaller in third year
- There are from 50 to 189 more children in unbalanced analysis

V-C and Correlations HR – Preferred Models

Type = UN: Balanced data

1	2	3	4
78.6	49.1	25.2	41.9
.60	85.6	42.0	59.2
.36	.58	62.1	51.9
.48	.65	.67	98.1

Type = UN: Unbalanced data

1	2	3	4
105.7	56.8	30.4	44.0
.56	98.0	43.7	54.5
.36	.54	68.0	53.2
.45	.58	.68	90.1

- Correlation higher between year 3 and 4
- As children age there may be higher correlation between adjacent years

Implications

- Inferences for fixed effects may be impacted by the poor choice of covariance matrix
- However, if

Number of subjects

- Number of treatment groups

- Number of repeated measures

$\geq$ 30

then UN will usually suffice

Implications

- MV GLM results are **identical** to PROC MIXED results when ddfm=kenwardrogers is used

Conclusions

- Univariate GLM V-C calculations are **identical** to MIXED estimates using type=CS for balanced data sets (ICC)
- MV GLM V-C estimates are **identical** to MIXED estimates using type=UN
- If sphericity assumption violated, MV GLM cannot determine best fit

Conclusions

- The ability to model a broader class of V-C structures yield results that make more sense in the context of the problem
- WT variability increases with age
- Correlation higher between adjacent years

Conclusions

- ARH(1) appears useful in explaining how WT tracks in very young children
- The UN nature of the relationship between HR measurements aids in assessment of the quality of this measurement

Conclusions

- A common correlation (ICC) may not be useful as an estimate of the relationship between repeated measurements
- PROC MIXED gives an opportunity to understand and quantify the true inter-relationships between repeated measurements

Contact information

- Maribeth Johnson
- Department of Biostatistics, AE-3035
- Medical College of Georgia
- Augusta, GA 30912-4900
- 706-721-0813
- majohnso@mcg.edu