

Intraclass Correlation Reliability Estimation Using the GLM and MIXED Procedures.

Maribeth Johnson
Office of Biostatistics
Medical College of Georgia, Augusta, GA

ABSTRACT

Intraclass correlation (ICC) reliability is used to assess the magnitude of measurement free from error found when using repeated measurements on individuals with the same or comparable instruments. For balanced data, ICC is defined as the ratio of mean covariances to mean variances. The estimation of the components of variation is therefore important to the estimation of reliability. This paper compares the variance component estimates obtained from the GLM and MIXED procedures of SAS/STAT® on two sets of data, one of which has missing data. ICC reliabilities based on the variance component estimates are also shown.

INTRODUCTION

The error involved in repeating the same set of measurements is what is meant by unreliability. At the same time, however, repeated measurements will ordinarily show some consistency. This tendency toward consistency from one set of measurements to another is called reliability. The degree of reliability of a set of measurements is an important consideration in research and for the future implementation of these measures in practice. In any study of prediction it is desirable to have a high degree of reliability of the measurement that is being predicted. One cannot make accurate prediction of an unreliable measurement.

DATA

The data used in the paper are from The Family Health Project that was conducted at the Texas site in the Studies of Child Activity and Nutrition (SCAN) program, a National Heart, Lung and Blood Institute funded 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 (Baranowski, et al. 1993). During the initial year of study the subjects included children aged 3 or 4 years. Families participated in four annual

summer clinics, held at the University of Texas, Galveston, at which time a variety of physiological and self report measures were collected. the first annual summer clinic was in 1986. Many children were seen at the first clinic and then dropped out of the study for various reasons.

Table 1. Data structure

<u>Number of records</u>	<u>Number of children</u>
1 of 4	102
2 of 4	34
3 of 4	36
4 of 4	98
	270
<u>Year</u>	<u>Number of observations</u>
1986	243
1987	148
1988	144
1989	135

The variables used in these analyses are weight (WT) and sum of seven skinfolds (SUM7). WT (nearest 0.1 kg) was measured using a Detecto balance-beam scale. Skinfold sites (mm) were biceps, triceps, subscapular, suprailiac, abdomen, thigh, and calf. Three measurements were taken in rotation and mean values were computed. The SUM7 is used as an indicator of adiposity. WT is generally thought of as a mechanized measurement that is relatively error free, while collecting skinfold measurements requires more lab technician skill and thus is subject to more measurement error.

USING PROC GLM

In the past a mixed model analysis of variance has been used to estimate reliability of measurements. For these data this was accomplished in PROC GLM using the statements:

```
proc glm; class child year;
model wt sum7 = child year;
random child;
```

The random statement outputs the expected mean

square (MS) for the between child variation of the form

$$\text{Var(Error)} + k \text{Var(CHILD)}$$

where the value of k is the average number of observations per child which is equal to the number of years for the balanced data. CHILD is treated as a fixed effect as far as the model fit is concerned, however. Since it is necessary to adjust for differences between years the Error MS is used as the Var(Error) , i.e. the within child variation, in the calculation. Therefore the between child variation and ICC are calculated as,

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

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

This ICC is the reliability of a single year of measurement. While this estimator of reliability is biased the individual components of variation are unbiased maximum-likelihood estimators when the data are balanced. See Winer (1971) for an explanation of how to obtain an unbiased estimator of the ICC from the mean square estimates.

This mixed model assumes that the error of measurement has a constant variance and that the additive random variable for child is independent of the measurement from year to year. This assumption implies that the measurement correlation is constant between any two years and that the total variance of a measurement is the same for each year. In the notation of multivariate analysis, the variance-covariance matrix of a child's vector of responses over time is compound symmetric. Thus, if the variability of a measure changes from year to year, as happens with certain growth parameters in young children, this model is inappropriate. Additionally, the assumption of equal correlation between any two years may not be valid. Importantly, when any of the assumptions of compound symmetry do not hold, the simple expression of ICC is also not useful as an overall definition of measurement reliability. More general models are needed, particularly those which provide for more general structures of variance-covariance relationships.

USING PROC MIXED

The ability to model irregular changes over time

using different covariance structures available in PROC MIXED may increase our understanding of the reliability of these measurements. The basic model that was used is as follows:

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

where *depvar* was either WT or SUM7. The REPEATED statement models the covariance structures in R, the variance-covariance matrix of the vector of errors. The SUBJECT=CHILD option is the mechanism for block diagonalizing R since subjects are considered independent. Four different *cov-structures* were considered, these are:

- 1) Compound symmetric (CS): This structure is the assumption of the ANOVA estimates of the variance components; therefore the results for the balanced data should be identical to the estimates from this analysis.
- 2) Heterogeneous compound symmetric (CSH): This structure assumes a common correlation between years but allows for different variances along the diagonal.
- 3) Heterogeneous first-order autoregressive (ARH(1)): This structure also allows for different variance parameters and produces an estimate of the autoregressive parameter so that the correlations between years separated by the same amount of time are the same, and the correlations of different amounts of time are also structured in the usual way, i.e. the correlation is ρ^m where m is the number of years between measures.
- 4) Unstructured (UN): This structure produces estimates of all four variances and six covariances in each subject block of R. Therefore, all of the correlations between years may be different. For balanced data, each block of the R matrix includes the sample variances and covariances of the dependent variable. (This is not the case, however, if the data are unbalanced.)

Likelihood ratio tests (LRT) were used to compare models for which one is a special case of the other (Wolfinger, 1992). All analyses were performed on a balanced data set containing 98 children, and the unbalanced data set containing observations on all

270 children that ever came to the clinics.

RESULTS AND DISCUSSION

The variances, covariances and correlations from the R and RCORR options of the REPEATED statement for WT using the balanced data set are shown in Table 2. All parameters estimated from the CS structure are the same as those calculated using the GLM mean square estimates. The between child variance is equal to 8.86, which is the covariance between years in the MIXED setup. The within child variance is 1.90, which when added to the between child variance is the same as the year variance of 10.76 from MIXED. Therefore, the reliability of one measurement is 0.823 from both analyses that assume equal year variances and common correlations between years. When the constraint of equality of variances between years is relaxed using the CSH structure, the common correlation between years, and therefore the reliability of one measurement, is estimated to be 0.92. When the common correlation constraint is structured using ARH(1), the reliability of measurements separated by one year is 0.951, by two years is 0.905 and by three years is 0.861. The UN structure produced very similar results. The LRT between these two models did not show a significant difference in fit, as seen in Table 4.

Table 2. Covariances and Correlations for WT using Different Covariance Structures, Balanced Data.

Type = CS:				
<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	
10.76	8.86	8.86	8.86	
1.0	.823	.823	.823	
(Variance and covariances in top line, correlations in bottom line)				
Type = CSH:				
<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	
4.71	5.04	6.61	9.21	
.920	6.37	7.69	10.71	
.920	.920	10.96	14.05	
.920	.920	.920	21.27	
(Variances on diagonal, covariances above, correlations below)				
Type = ARH(1):				
<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	
4.49	5.12	6.44	8.44	
.951	6.45	8.12	10.63	
.905	.951	11.30	14.78	
.861	.905	.951	21.37	
(Variances on diagonal, covariances above, correlations below)				
Type = UN:				
<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	
4.61	5.28	6.45	8.39	
.961	6.55	8.12	10.61	
.900	.951	11.13	14.32	
.858	.910	.942	20.77	
(Variances on diagonal, covariances above, correlations below)				

The ARH(1) covariance structure appears to be the best fit for these data, so the assumption of a common correlation between years does not hold and neither does the idea of a single reliability estimate to describe the relationship of subject and measurement error. What has been done in the past does not adequately describe the reliability of this measurement.

The results of the WT analyses from the unbalanced data set are shown in Table 3. With unbalanced data the ANOVA estimates from GLM are no longer maximum-likelihood estimators so the parameter estimates are not identical to the CS structure from PROC MIXED, although they are similar. The between child variance is estimated to be 9.95 from GLM and 9.17 from MIXED. The within child (error) variances are 2.03 and 2.00 from GLM and MIXED, respectively. These result in slightly different reliability estimates. The remaining results are similar to the balanced case.

Table 3. Covariances and Correlations for WT using different Covariance Structures, Unbalanced Data.

Type = CS:				
<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	
11.17	9.17	9.17	9.17	
1.0	.820	.820	.820	
(Variance and covariances in top line, correlations in bottom line)				
Type = CSH:				
<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	
16.11	6.80	9.37	12.16	
.937	.860	11.12	14.42	
.937	.937	16.37	19.82	
.937	.937	.937	27.53	
(Variance on diagonal, covariances above, correlations below)				
Type = ARH(1):				
<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	
5.97	7.02	9.24	11.79	
.962	8.94	11.77	15.01	
.925	.962	16.75	21.36	
.889	.925	.962	29.45	
(Variances on diagonal, covariances above, correlations below)				
Type = UN:				
<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	
5.97	7.06	8.99	11.52	
.966	8.94	11.61	14.98	
.904	.955	16.54	21.21	
.871	.925	.964	29.30	
(Variances on diagonal, covariances above, correlations below)				

By relaxing the constraint of variance equality using the CSH structure the model is seen to have a better fit over CS (Table 4). The same holds true when both the variance and correlation equality constraints are relaxed using the ARH(1) structure. As was the case using balanced data, the fit from

the UN model was not significantly better than that from ARH(1). The interesting finding is that in all cases the parameter estimates are very similar to those estimated from the balanced data even though there are anywhere from 37 to 145 more children measured in any one year.

Table 4. REML Likelihood Ratio Tests between Covariance Structures for WT.

Balanced Data:

Type	Parameters	CM	df	-2LL	χ^2
CS	2	-	-	1658	-
CSH	5	CS	3	1370	288**
ARH(1)	5	CS	3	1292	366**
UN	10	CSH	5	1284	86**
UN	10	ARH(1)	5	1284	8

Unbalanced Data:

Type	Parameters	CM	df	-2LL	χ^2
CS	2	-	-	3012	-
CSH	5	CS	3	2601	411**
ARH(1)	5	CS	3	2505	507**
UN	10	CSH	5	2496	105**
UN	10	ARH(1)	5	2496	9

CM is the comparison model for the likelihood ratio test

**p<.0005

The analyses for SUM7 are included to assess the impact of a measurement that has more error associated with it. For the balanced data (Table 5) the ANOVA estimates and those from the CS structure are again identical. The reliability estimate from both analyses is 0.757. This reliability increases to 0.853 when the variances are allowed to be heterogeneous under the CSH structure. When a first-order autocorrelation structure with heterogeneous variances is fit using ARH(1) the reliability of measurements separated by one year is 0.895, by two years is 0.80 and by three years is 0.716. The convergence criterion had to be relaxed to 0.0001 in order for the ARH(1) model to converge. It reached this criterion after 2 rounds of iteration but then remained there for 48 rounds. The variances estimated by both the ARH(1) and UN structures are identical. However, the problem of failure to converge arises when a first-order autocorrelation is fit to data in which the correlations across time are not AR(1). In this data, the correlation between successive years in the UN analysis is similar to the correlation in the ARH(1) model, but the correlations between years that are more separated do not decrease as quickly across time in the UN analysis as in the ARH(1) model. However, the fit is not significantly better for UN

than the ARH(1) structure, as seen in the LRT shown in Table 7.

Table 5. Covariances and Correlations for SUM7 using Different Covariance Structures, Balanced Data.

Type = CS:

<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>
516.8	391.1	391.1	391.1
1.0	.757	.757	.757

(Variances and covariances in top line, correlations in bottom line)

Type = CSH:

<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>
198.4	209.3	286.2	381.8
.853	303.8	354.2	472.4
.853	.853	568.1	645.9
.853	.853	.853	1,010.5

(Variances on diagonal, covariances above, correlations below)

Type = ARH(1):

<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>
192.4	220.0	268.5	310.1
.895	314.3	383.6	443.1
.800	.895	585.1	675.8
.716	.800	.895	975.4

(Variances on diagonal, covariances above, correlations below)

Type = UN:

<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>
192.4	219.3	280.9	327.7
.892	314.3	389.5	463.6
.837	.908	585.1	665.8
.756	.837	.881	975.4

(Variances on diagonal, covariances above, correlations below)

The variance and covariance estimates of SUM7 using the unbalanced data (Table 6) closely match those from the balanced data analyses. The convergence criterion or the ARH(1) structure had to be relaxed to 0.005 in order to obtain convergence and it reached this in the first round, remaining flat over the next 49 rounds.

Table 6. Covariances and Correlations for SUM7 using Different Covariance Structures, Unbalanced Data.

Type = CS:

<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>
504.8	371.7	371.7	371.7
1.0	.736	.736	.736

(Variance and covariances in top line, correlations in bottom line)

Type = CSH:

<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>
219.7	249.1	355.0	429.7
.861	381.1	476.6	566.0
.861	.861	774.1	806.7
.861	.861	.861	1,134.1

(Variances on diagonal, covariances above, correlations below)

Type = ARH(1):

1	2	3	4
205.7	251.5	336.1	336.4
.899	380.6	508.7	509.1
.808	.899	841.5	842.3
.726	.808	.899	1,043.5

(Variances on diagonal, covariances above, correlations below)

Type = UN:

1	2	3	4
213.4	257.7	338.0	379.5
.887	395.8	515.8	582.0
.817	.916	801.6	880.1
.758	.853	.907	1,175.1

(Variances on diagonal, covariances above, correlations below)

For the unbalanced data, the model containing the UN structure had a significantly better fit ($p < .05$) than that containing the ARH(1) structure (Table 7) although the fits were very close. The first-order autoregressive structure with heterogeneous variance may not adequately describe the structure of SUM7 as a measurement of adiposity in children, but it does a better job than the CS assumption of the past.

Table 7. REML Likelihood Ratio Tests between covariance Structures for SUM7.

Balanced Data:

Type	Parameters	CM	df	-2LL	χ^2
CS	2	-	-	3247	-
CSH	5	CS	3	3040	207**
ARH(1)	5	CS	3	3007	240**
UN	10	CSH	5	2999	41**
UN	10	ARH(1)	5	2999	7

Unbalanced Data:

Type	Parameters	CM	df	-2LL	χ^2
CS	2	-	-	5691	-
CSH	5	CS	3	5380	311**
ARH(1)	5	CS	3	5337	354**
UN	10	CSH	5	5325	55**
UN	10	ARH(1)	5	5325	12*

CM is the comparison model for the likelihood ratio test

** $p < .0005$, * $p < .05$

CONCLUSION

The results given by the more general variance component models yield results which seem to make more sense in the context of the problem. We would expect the variability of weight in children to increase from age 3-4 to age 6-7, and for the correlation of weight in adjacent years to be higher than the correlation of weight among years that are

more separated. These findings impact how researchers plan future studies.

In a sampling situation in which there is subject and measurement variability, the components of variance are used to determine how many measurements are needed to achieve a certain level of reliability. This level is chosen to be high enough so that accurate predictions and inferences can be modeled from these measurements. When the variances and covariances are changing over time, the sample size determination requires more thought. When the research objective is to use changes of a dependent variable over time as a predictor variable, the specific years of change for the prediction will influence the sample size required, unlike the previous approaches taken to this problem.

The unusually large values of -2LL for the SUM7 data suggest that while the UN model provides a better fit than models previously examined, there may be a problem with the model specification. Additional exploration of this data set is warranted.

REFERENCES

- Baranowski T, Stone ET, Klesges RK, et al. (1993) "Studies of child activity and nutrition (SCAN): Longitudinal research on CVD risk factors and CVH behaviors in your children." *Cardiovascular Risk Factors*, 2, 4-16.
- SAS Institute Inc. (1989), *SAS/STAT User's Guide: Version 6, Fourth Edition, Volume 2*, Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (1992), *SAS Technical Report P-229, SAS/STAT Software: Changes and Enhancements, Release 6.07*, Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (1994), *SAS/STAT Software: Changes and Enhancements, Release 6.10*, Cary, NC: SAS Institute Inc.
- Winer, BJ (1971), *Statistical Principles in Experimental Design*, Second Edition, New York: McGraw-Hill, Inc.
- Wofinger, RD (1992), "A tutorial on mixed models", Cary, NC: SAS Institute, Inc.
- SAS and SAS/STAT are registered trademarks or trademarks of SAS Institute Inc. In the USA and other countries. ® indicates USA registration.