Revista Electrónica de Metodología Aplicada
2007, Vol. 12 nº 2, pp. 15-31

Using linear mixed models in longitudinal studies: Application of SAS PROC MIXED*

Roser Bono[1], Jaume Arnau[1] y Nekane Balluerka[2]

[1]Department of Methodology of the Behavioural Sciences, Faculty of Psychology, University of Barcelona.
[2]Department of Social Psychology and Methodology of the Behavioural Sciences, University of the Basque Country.

ABSTRACT

The objectives of this article are twofold: (a) to outline the basic concepts associated with the linear mixed model and (b) to illustrate how this model can be used to analyse systematic interindividual differences in intraindividual change, this being achieved through a longitudinal study of a cohort of children living in Cordoba (Argentina). These objectives will be met by using the PROC MIXED statement of the SAS software. This software fits a wide variety of linear mixed models to longitudinal data, thus enabling valid statistical inferences to be made. Since the choice of covariance structure may influence the values obtained in significance tests for fixed effects, we focus our attention on this aspect. The most common covariance structures for modelling longitudinal data are described and guidelines are proposed for choosing the structure which enables more powerful and more efficient regression parameter estimates to be made.

Keywords: longitudinal data, repeated measurements, linear mixed model, multilevel modelling, covariance structures, PROC MIXED.

* Acknowledgements

1. We are grateful to the CLACYD Foundation for authorising, via an agreement between the University of Barcelona and CLACYD, the use of the data analysed in this article.

2. The research was supported by Grant SEJ2005-01923/PSIC under the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica of Spain's Ministerio de Educación y Ciencia.

Corresponding author:
Roser Bono
Facultad de Psicología. Universidad de Barcelona.
Departamento de Metodología de las Ciencias del Comportamiento.
C/ Passeig de la Vall d'Hebrón, 171
08035 BARCELONA
SPAIN
Phone: 34-3-312 50 80 Fax: 34-3-402 13 59
e-mail: rbono@ub.edu

1.- Introduction

Over the last decade, social and health researchers have turned their attention to the relationships between individuals and the social milieu in which they develop, the aim being to evaluate the influence of context on individual behaviour. This has particular relevance in the field of education where, in addition to the individual progress of students, it is important to analyse the influence of schools themselves. Likewise, there is a long tradition within health research of studies analysing the variability between individuals from different geographical areas or different groups. This type of study provides information about the health of individuals and the area in which they live, or about the treatment received by patients and the characteristics of the health centre they attend. Therefore, different hierarchies of available information are established. In most situations only two hierarchical levels are considered, although three or more may be studied. For example, within the area of health, patients (level 1 unit) are grouped into hospitals (level 2), which in turn are grouped into geographical areas (level 3). In education an example of a three-level hierarchical structure would be pupils in classrooms within schools.

Establishing this hierarchy of different variables has important repercussions for the data analysis. It is assumed that subjects belonging to the same group will tend to be more similar to one another than to members of other groups, and this similarity between individuals yields an intra-group correlation structure that rules out the use of traditional estimation methods. Therefore, considerable efforts have been made to analyse these hierarchical structures via approaches that enable valid statistical inferences to be made. The result of this is what are termed multilevel models, based on the linear mixed model.

The hierarchical structure can also be applied to situations where repeated measures are taken of subjects, that is, in longitudinal studies. In the longitudinal field, a typical two-level hierarchical structure would distinguish, on the first level, the repeated observations per subject and, on the second level, the individuals themselves. According to this hierarchical structure the individual effects and intraindividual variation, which is a function of time, define the first level, while the interindividual variation associated with the subjects' characteristics define the second level. As longitudinal studies are usually characterised by autocorrelation between the observations of the same subject and by the sample attrition, they cannot be analysed by means of traditional regression models. Thus, multilevel models are currently used instead of these traditional models. Various authors have argued that multilevel models are the most suitable for the analysis of longitudinal data (Bock, 1989; Bryk & Raudenbush, 1992; Goldstein, 2003; Hoeksma & Knol, 2001; Plewis, 2001; Raudenbush, 1989; Snijders, 1996). Using this analytic procedure it is possible to determine individual growth profiles and infer the effect of the variables that produce variance between subjects (Hox, 2002).

The main aim of data analysis using the linear mixed model is to define an adequate error covariance structure in order to obtain efficient estimates of the regression parameters. The statistical software now includes the covariance structure as part of the statistical model and thus the covariance matrix can be used to estimate the fixed effects of treatment and time by means of the generalized least squares method. Noteworthy among this statistical software are the HLM (Bryk & Raudenbush, 1992; Bryk, Raudenbush, & Congdon, 1996; Raudenbush, Bryk, Cheong, & Congdon, 2000), the MLwiN (Prosser, Rasbash, & Goldstein,

1996; Rasbash et al., 2000) and the SAS (Littell, Milliken, Stroup, & Wolfinger, 1996; Sheu & Suzuki, 2001; Verbeke & Molenberghs, 1997). The present study, in the form of a tutorial, describes how to analyse longitudinal data using the PROC MIXED of the SAS system (SAS Institute Inc, 2000, 2004). The first part provides a brief description of the linear mixed model, whose application is then illustrated by studying data concerning the weight of a group of children from birth to five years of age. This study will serve to demonstrate how to model the error covariance structure which, undoubtedly, is the core of the linear mixed model.

2.- Linear mixed model

The linear mixed model is an application of the general linear model and is defined in the following terms:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}_i + \mathbf{e}_i \quad (1)$$

where $\mathbf{y}_i$ is the vector of repeated measures data for the i-th subject, $\mathbf{X}_i$ is the known design matrix that includes covariables for the fixed effects, $\boldsymbol{\beta}$ is the vector of fixed-effects parameters, $\mathbf{Z}_i$ is another known design matrix that includes covariances for the random effects, $\boldsymbol{\gamma}_i$ is the vector of random-effects parameters or the residuals at the subject level and $\mathbf{e}_i$ is the vector of the errors at the level of observation. The fixed part of the model is specified by $\mathbf{X}_i\boldsymbol{\beta}$ and the random part by $\mathbf{Z}_i\boldsymbol{\gamma}_i + \mathbf{e}_i$.

The linear mixed model does not impose conditions on the covariance structures; it merely assumes that $\boldsymbol{\gamma}_i$ and $\mathbf{e}_i$ have an independent multivariate normal distribution with a zero mean and covariance matrices $\mathbf{G}$ and $\mathbf{R}_i$, respectively. On the basis of this model, the observations vector $\mathbf{y}_i$ has a multivariate normal distribution with an expected value of

$$E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta} \quad (2)$$

and variance

$$V(\mathbf{y}_i) = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i \quad (3)$$

where $\mathbf{Z}_i\mathbf{G}\mathbf{Z}_i'$ is the between-subjects component and $\mathbf{R}_i$ the within-subject component, such that the covariance matrix of observations is a function of $\mathbf{G}$ and $\mathbf{R}_i$, where $\mathbf{G}=V(\boldsymbol{\gamma}_i)$ and $\mathbf{R}_i=V(\mathbf{e}_i)$. Note that the fixed effects define the expected values of the observations, while the random effects represent the variances and covariances of the observations (Littell, Pendergast, & Natarajan, 2000).

In longitudinal data, the repeated measures can be considered as dependent variables. For that reason, the multivariate analysis is a good alternative to the univariate analysis. However, the advantage of the linear mixed model over traditional analytic approaches to longitudinal data is that it models the covariance matrix. Thus, the fixed parameter estimates are more efficient and the model is more powerful in terms of testing the effects associated with the repeated measures. This approach is also more robust than traditional univariate and multivariate tests. However, when the covariance structure is not adequately fitted and sample sizes are small, a positive bias in type I error is produced (Wright & Wolfinger, 1997).

3.- Analysis of an empirical study: Growth differences in children

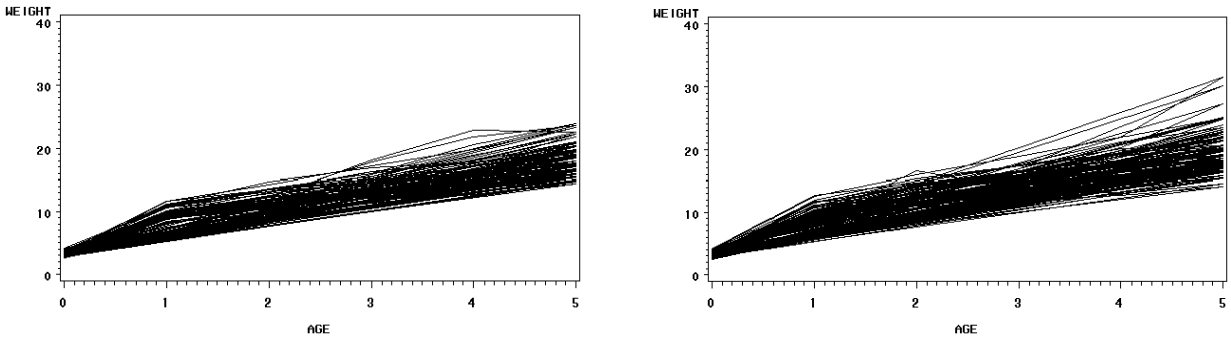
This section describes the use of the PROC MIXED in the SAS system (version 9.1.3) with data concerning the weight of a group of children from birth to five years of age. These data form part of a study conducted by the CLACYD Foundation (the initials in Spanish standing for Cordoba, breast-feeding, food, growth and development). The research began in 1993 and the cohort was followed up over the five-year period to 1998. All births occurring between 10 and 22 May 1993 in public and private institutions offering obstetric services were recorded. The inclusion criteria were that the babies had to live in Cordoba (Argentina), have a minimum weight of 2.5 kilos, not be the result of a multiple pregnancy and be free of any malformations (Sabulsky et al., 2001). As it is an illustrative example we used a sub-sample of 140 subjects (65 boys and 75 girls). As the dependent variable, weight in kilos was measured at birth and on five subsequent occasions (at 1, 2, 3, 4 and 5 years). The aim of the study was to examine whether there were systematic interindividual differences in intraindividual change in the children's weight over time as a consequence of the method of feeding (breast *vs*. bottle). The breast-fed group, comprising 56 subjects (25 boys and 31 girls), were fed solely breast milk during the first four months of life. The bottle-fed group, consisting of 84 subjects (40 boys and 44 girls), were either never breast-fed or weaned during the first two months of life.

Table 1 shows the set of variables that form part of the study. The data file (weight.dat) has been organised with the SAS system following the repeated measures format (SAS Institute Inc, 2000), such that birth weight is taken as the baseline and the successive observations constitute the repeated measures.

Variable	Description
PERSON	Person in the study (140 levels)
WEIGHTBL	Baseline (birth weight)
FEEDING	Feeding method (breast *vs*. bottle)
WEIGHT	Weight across five-year period for each subject
AGE	Age from 1 year to 5 years (5 levels)

Table 1. Variables in the SAS Data Set Weight

Figure 1 shows the within-subject profile graphs according to the type of feeding during the first five years of life. Note that the profiles follow a similar pattern in both groups of children: marked increase in weight during the first year of life and a growing linear trend. It can also be seen that the between-subjects variation, particularly in the bottle-fed group, increases with age.



BREAST-FED GROUP BOTTLE-FED GROUP

Figure 1. Profile graphs according to feeding method (breast *vs*. bottle)

Figure 2 shows the mean weight values over time as a function of the method of feeding (breast *vs*. bottle). The patterns coincide with those of the profile graphs (Figure 1). During the first year there is a large increase in weight and there are no differences between the children according to the method of feeding. Both groups present similar means at AGE=0 or baseline and AGE=1. After year one the growth slope is less steep and a pattern of interindividual change appears as a function of the method of feeding. It can also be seen that the between-groups differences remain constant over time, the weight of bottle-fed children being greater.

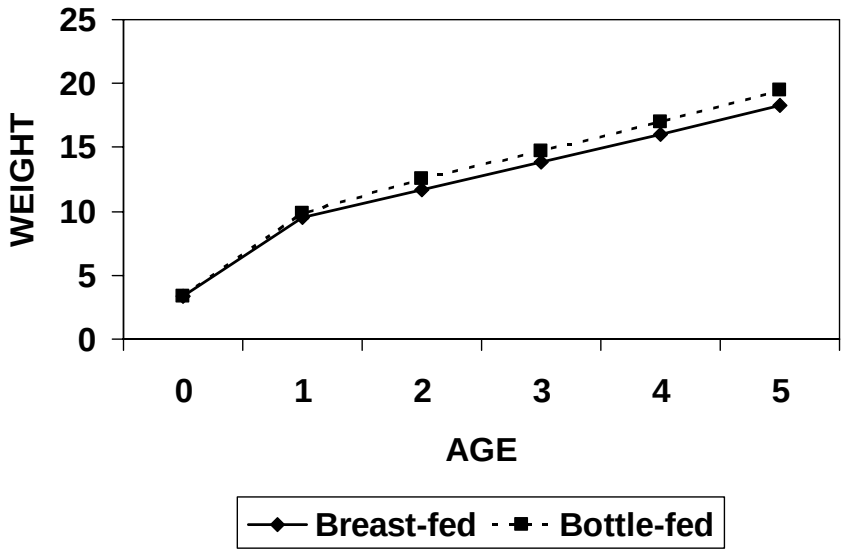


Figure 2. Mean weight according to feeding method (breast *vs*. bottle)

As pointed out above, the linear mixed model enables a given covariance structure to be defined in order to obtain efficient regression parameter estimates. In the next section we describe the four covariance structures whose fit was examined in the present study. It will be seen that different information criteria were used in each one of the models in order to select the structure that offered the best fit.

3.1.- Linear mixed models using SAS PROC MIXED

The PROC MIXED program of the SAS system is suitable for fitting mixed models. In order to apply this program it is necessary to specify the components of equation 1 and determine the covariance structure of $\mathbf{Z}_i\boldsymbol{\gamma}_i$ and $\mathbf{e}_i$. PROC MIXED can be used not only to estimate the fixed parameters of $\boldsymbol{\beta}$, but also the covariance parameters of $\mathbf{G}$ and $\mathbf{R}_i$. By default, PROC MIXED estimates the covariance parameters using the method of restricted maximum likelihood (REML), also known as residual maximum likelihood.

The mixed model is specified by means of the CLASS, MODEL, RANDOM and REPEATED statements. The CLASS statement identifies the classification variables (for example, gender, person, age, etc.). The MODEL statement specifies the model's fixed effects equation, $\mathbf{X}_i\boldsymbol{\beta}$. Thus, the design matrix $\mathbf{X}_i$ is defined and the model's intercept is included by default. The RANDOM statement contains the random effects, $\mathbf{Z}_i\boldsymbol{\gamma}_i$, including the structure of $\mathbf{G}=V(\boldsymbol{\gamma}_i)$. The REPEATED statement models the intraindividual variation and includes the structure of $\mathbf{R}_i=V(\mathbf{e}_i)$, where $\mathbf{R}_i$ is a block diagonal matrix for each subject. If the REPEATED statement is not included it is assumed that $\mathbf{R}_i=\sigma^2\mathbf{I}$.

To illustrate the analysis we compared the influence of the method of feeding (breast *vs*. bottle) on the evolution of children's weight from birth to five years of age and used the baseline WEIGHTBL as the covariable. Prior to carrying out the analysis, different structures of the matrix $\mathbf{R}_i$ were fitted. In addition to the simple model, which corresponds to an ANOVA, we also took into account a number of more frequently used models, such as the unstructured model, the compound symmetry model and the first-order autoregressive model. The most typical covariance structure for longitudinal data is the unstructured model, as it requires no assumption regarding the error terms and allows any correlation pattern between the observations. However, when it is assumed that the correlations between the observation points are constant, the covariance structure takes the compound symmetry form. This assumption is supported by the repeated measures ANOVA and is not very common with real data. Finally, a common structure in longitudinal data is the autoregressive one. This structure falls between the unstructured and compound symmetry models. In all these models it is assumed that each subject has the same covariance structure and that the data from different subjects are independent.

An additional feature of the PROC MIXED is that it allows the user to specify, separately and jointly, covariance structures that assume within-subjects and/or between-subjects heterogeneity. Within-subjects heterogeneity occurs when the variances across repeated measures are unequal while between-subjects heterogeneity occurs when covariance matrices differ across groups. In this study, we analyzed a model assuming within-subjects

heterogeneity because the data of the example are characterized by this structure, and we compared this model with models that assume within-subjects homogeneity.

3.1.1 Simple (VC)

The simple model assumes independent observations and homogeneous variance. Consequently, **G**=0 and $\mathbf{R}_i = \sigma^2 \mathbf{I}$.

$$VC = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (4)$$

This model cannot be used with repeated measures obtained from the same subject due to their non-independence. The syntax of the PROC MIXED, when it doesn't include the RANDOM or REPEATED statement, corresponds to the general linear model:

```
proc mixed data = weight covtest;
        class feeding person age;
        model weight = weightbl feeding age feeding*age;
```

The instruction DATA=*name of data file* allows access to the data set and the option COVTEST prints the statistical significance of the estimated covariance parameters. The CLASS statement includes the classification variables FEEDING, PERSON and AGE. In the MODEL statement, WEIGHT is defined as a response variable and WEIGHTBL represents the regression effect of the baseline. The variable FEEDING, due to its being a classification variable, models a different mean for each level of feeding method. Similarly, the different means are specified for the levels of AGE. Finally, the interaction FEEDING*AGE is modelled. Another way of expressing this same model is by adding the option TYPE=VC of the REPEATED statement. In this example, the variance estimate of the VC model is $\sigma^2 = 3.3088$.

3.1.2.- Unstructured (UN)

In the unstructured covariance matrix all the variances and covariances are different. Although this structure is the most heterogeneous it also offers the best fit. However, its application requires the estimate of many parameters, $K(K+1)/2$, where K is the number of repeated measures or observations. With the five repeated measures of the example analysed the matrix adopts the following pattern:

$$\text{UN} = \begin{bmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} & \sigma_{51} \\ \sigma_{21} & \sigma_2^2 & \sigma_{32} & \sigma_{42} & \sigma_{52} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{43} & \sigma_{53} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 & \sigma_{54} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_5^2 \end{bmatrix} \quad (5)$$

The PROC MIXED that fits the unstructured model is specified as:

```
proc mixed data = weight covtest;
        class feeding person age;
        model weight = weightbl feeding age feeding*age;
repeated/type = un subject = person r rcorr;
```

The FEEDING and PERSON variables included in the CLASS statement are classification variables. As the AGE variable contains only a few levels it can also be treated as a classification variable in order to avoid biased estimates of the variance and covariance parameters (Littell et al., 2000). The fixed effects, specified in the MODEL statement, correspond to WEIGHTBL, FEEDING, AGE and FEEDING*AGE. The unstructured covariance matrix (TYPE=UN) is specified using the REPEATED statement, in other words, in terms of $\mathbf{R}_i$ and assuming $\mathbf{G}$=0. The option SUBJECT=PERSON defines $\mathbf{R}_i$ as a block diagonal matrix with a sub-matrix for each individual. Finally, the options R and RCORR require the $\mathbf{R}_i$ matrix to be printed in terms of covariance and correlation, respectively.

The elements on the diagonal of Table 2 are the estimates of the between-subjects variances with respect to feeding method at different ages. Note that, as in Figure 1, the estimated variances increase over time, from $\sigma_1^2 = 0.8938$ to $\sigma_5^2 = 7.8834$. The covariances are situated above the diagonal and the correlations below it. The correlations between the values of the WEIGHT variable become weaker as the distance between observations increases. For example, the correlation of weights between AGE=1 and AGE=2 is 0.7632, while that between AGE=1 and AGE=5 is 0.6106. This indicates a trend towards weaker correlations as the interval between measures increases. With this type of data an ANOVA is not suitable.

AGE 1	AGE 2	AGE 3	AGE 4	AGE 5
0.8938	0.8921	1.0098	1.2798	1.6207
0.7632	1.5286	1.4229	1.7209	2.1504
0.7163	0.7718	2.2236	2.8210	3.3660
0.6600	0.6790	0.9229	4.2017	5.2374
0.6106	0.6195	0.8040	0.9100	7.8834

Note: Variances on diagonal, covariances above diagonal, correlations below diagonal

Table 2. Covariance and Correlation Estimates for the Unstructured Matrix

3.1.3.- Compound Symmetry (CS)

The compound symmetry structure assumes that the observations of the same subject have homogeneous variances and homogeneous covariances. With five repeated measures the matrix adopts the following pattern:

$$CS = \begin{bmatrix} \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 \end{bmatrix} \quad (6)$$

The covariance matrix CS can be specified with PROC MIXED in two different ways: with the RANDOM statement or with the REPEATED statement. However, the latter procedure is preferable as it can also be used when the within-subject correlation is negative.

With the REPEATED statement and the options TYPE=CS and SUBJECT=PERSON, the matrix $\mathbf{R}_i$ is defined with two unknown parameters, one which models the common covariance and the other the residual variance:

$$\mathbf{R}_i = \begin{bmatrix} \sigma_1^2 + \sigma^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma^2 & \sigma_1 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma^2 \end{bmatrix} \quad (7)$$

The PROC MIXED instructions specify the structure for each individual subject and print the sub-matrix $\mathbf{R}_i$ of a subject in terms of covariance and correlation (R and RCORR, respectively):

```
proc mixed data = weight covtest;
        class feeding person age;
        model weight = weightbl feeding age feeding*age;
        repeated/type = cs subject = person r rcorr;
```

The parameter estimates of the matrix $\mathbf{R}_i$ obtained via the CS model are $\sigma_1^2 = 2.1265$ and $\sigma^2 = 1.1942$. According to this compound symmetry structure the estimated variance $\sigma_1^2 + \sigma^2 = 3.3207$ is similar to that obtained with the VC structure, since it is a special case of the simple model. The CS covariance structure supports the assumption of the repeated measures ANOVA. This assumption is restrictive and is not usually met by real data.

3.1.4.- First-order autoregressive (AR(1))

The first-order autoregressive model assumes that measurements which are close to one another in time will show high correlations. Its structure is homogeneous, the variances are equal and the covariances between observations of the same subject decrease exponentially as the lag increases. Therefore, it comprises two parameters: the parameter of the variance of observations and that of the correlation between adjacent observations:

$$\text{AR(1)} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} \quad (8)$$

The correlation ρ is that between the observations of interval one and two, while correlation ρ^2 is that between the observations of interval one and three, and so on. Therefore, the AR(1) correlations follow an exponential function:

$$\text{corr}_{\text{AR(1)}}(\text{lag}) = \rho_{\text{AR(1)}}^{\text{lag}} \quad (9)$$

The autoregressive structure is defined in terms of $\mathbf{R}_i$ and $\mathbf{G}$=0. This covariance structure is specified for each subject with the REPEATED statement:

```
proc mixed data = weight covtest;
        class feeding person age;
        model weight = weightbl feeding age feeding*age;
        repeated /type=ar(1) subject=person r rcorr;
```

The variance parameter estimated by the AR(1) model is 3.8943 and ρ = 0.8716. According to equation (9), for observations separated by one year the correlation is 0.8716; for observations separated by two years the correlation is 0.7597; for observations separated by three years the correlation is 0.6622; and for observations separated by four years the correlation is 0.5772. These values are considerably higher than the zero correlation assumed by the ordinary least-squares analysis.

Table 3 summarises the covariance and correlation matrices resulting from the simple, compound symmetry and autoregressive models. The unstructured covariance and correlation matrices are shown in Table 2.

	AGE 1	AGE 2	AGE 3	AGE 4	AGE 5
Simple (VC)	3.3084	0	0	0	0
	1	0	0	0	0
Compound Symmetry (CS)	3.3207	2.1265	2.1265	2.1265	2.1265
	1	0.6404	0.6404	0.6404	0.6404
Autoregressive (AR(1))	3.8943	3.3944	2.9586	2.5788	2.2477
	1	0.8716	0.7597	0.6622	0.5772

Note: Variances and covariances in top line, correlations in bottom line

Table 3. Variance, Covariance and Correlation Estimates for Covariance Structures

3.2.- Comparison of fits of covariance structures

With the mixed model it is possible to select the covariance structure which best describes the data. The PROC MIXED uses three fit criteria: -2 times the residual log-likelihood (-2RLL), Akaike's Information Criterion (AIC) (Akaike, 1974) or its corrected version for finite samples (AICC) (Hurvich & Tsai, 1989), and the Bayesian Information Criterion (BIC) (Schwarz, 1978). These criteria are indices of relative goodness-of-fit and may be used to compare models with different covariance structures and the same fixed effects (Bozdogan, 1987; Keselman, Algina, Kowalchuk, & Wolfinger, 1998; Littell et al., 1996; Wolfinger, 1993, 1996, 1997).

The interpretation of results must take into account, firstly, the null model likelihood ratio test. This test is statistically significant in well-fitted models, which indicates that they are better than the ordinary least-squares null model (H_0: $\mathbf{R}_i=\sigma^2\mathbf{I}$). It can be seen in Table 4 that the null model likelihood ratio test is highly significant ($p<0.0001$) for CS, AR(1) and UN covariance structures, and thus all the structures used in this example provide a better fit than the null model. The -2RLL criterion, which measures the deviance between the data and the model, can be used to determine which covariance structure is the most adequate. The smaller this index is, the better the fit. In this study the best fit corresponds to the unstructured covariance matrix. However, the -2RLL criterion cannot be directly interpreted, but only in comparison with other models (Singer, 2002; Singer & Willett, 2003; Luke, 2004). The difference of deviances between two models is distributed as a chi-squared with as many degrees of freedom as the difference between the number of estimated parameters in each model. In Table 4 the deviance of the UN model is 1857.5, while that of the CS and AR(1) models is greater. The difference between the deviances of the UN and CS model is 581.1, while that between the UN and AR(1) models is 292.9. These differences, when compared with a chi-squared distribution with 13 degrees of freedom (15 parameters – 2 parameters), are statistically significant ($p<0.001$). The UN model clearly shows the best fit. However, this model is not parsimonious; that is to say, it is unable to explain the data with few parameters. Therefore, the AIC and BIC fit criteria have been proposed on the basis of the deviance but they penalize due to the number of parameters to be estimated:

$$AIC = -2RLL + 2d \quad (10)$$

$$BIC = -2RLL + d\ln(N) \quad (11)$$

where d is the number of parameters to be estimated and N is the sample size.

As with the -2RLL criterion, the smaller the value of the AIC and BIC criteria, the better the fit. Examination of the AIC and BIC criteria confirms that the unstructured covariance matrix is the one which best fits the data (Table 4).

Structure Name	CS	AR(1)	UN
Covariance Parameters	2	2	15
-2RLL	2438.60	2150.40	1857.50
AIC	2442.60	2154.40	1887.50
BIC	2448.50	2160.30	1931.60
Chi-Square	387.94	676.17	969.12
Pr>Chi-Square	<.0001	<.0001	<.0001

Table 4. Fit Statistics and Null Model Likelihood Ratio Test

3.3.- Tests of fixed effects

Once the covariance structure has been selected the results from the tests of fixed effects can be interpreted. On the basis of the results obtained it can be concluded that all the fixed effects are statistically significant (WEIGHTBL, FEEDING, AGE, FEEDING*AGE).

Table 5 shows the statistical significance of the fixed effects for models with different covariance structures. This enables us to see how a given covariance structure affects the accuracy of the inference of fixed effects. It can be seen that although the interaction FEEDING*AGE of the VC model is not statistically significant (p=0.3817), it is significant in the other models. Furthermore, although the p values of the CS, AR(1) and UN models are statistically significant for all the fixed effects, the F values differ.

Structure name	WEIGHTBL	FEEDING	AGE	FEEDING*AGE
VC	59.88	35.07	521.56	1.05
	p<.0001	p<.0001	p<.0001	p=0.3817
CS	16.75	9.81	1444.90	2.90
	p<.0001	p=0.0021	p<.0001	p=0.0214
AR(1)	15.20	7.33	857.37	3.33
	p=0.0002	p=0.0076	p<.0001	p=0.0104
UN	20.93	9.67	749.50	4.73
	p<.0001	p=0.0023	p<.0001	p=0013

Table 5. Values of F Test for Fixed Effects for different Covariance Structures

The study of an interaction that is statistically significant can be completed with the LSMEANS statement:

```
lsmeans feeding*age /cl;
```

This statement enables us to obtain the means corresponding to the FEEDING*AGE combinations and, with the CL option, the confidence limits. The results of the LSMEANS statement for the UN model are shown in Table 6. When analysing the means of the

FEEDING*AGE interaction it can be seen that children in the bottle-fed group (FEEDING 2) are heavier, this effect being greater at four and five years of age.

Least Squares Means

Effect	Feeding	Age	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
feeding*age	1	1	9.5263	0.1263	137	75.41	<.0001	0.05	9.2765	9.7761
feeding*age	1	2	11.7255	0.1652	137	70.97	<.0001	0.05	11.3988	12.0522
feeding*age	1	3	13.8214	0.1993	137	69.36	<.0001	0.05	13.4273	14.2154
feeding*age	1	4	16.0307	0.2739	137	58.52	<.0001	0.05	15.4890	16.5723
feeding*age	1	5	18.2432	0.3752	137	48.62	<.0001	0.05	17.5013	18.9852
feeding*age	2	1	9.8353	0.1032	137	95.35	<.0001	0.05	9.6313	10.0392
feeding*age	2	2	12.5712	0.1349	137	93.19	<.0001	0.05	12.3045	12.8380
feeding*age	2	3	14.7095	0.1627	137	90.41	<.0001	0.05	14.3878	15.0312
feeding*age	2	4	16.9448	0.2237	137	75.76	<.0001	0.05	16.5025	17.3870
feeding*age	2	5	19.4346	0.3063	137	63.44	<.0001	0.05	18.8289	20.0404

Table 6. Least squares means output for the FEEDING*AGE combinations

Finally, if we want to estimate the effect of feeding method for a given individual, for example, one with a baseline (WEIGHTBL) value of 3.7, it is necessary to specify in the LSMEANS statement the following options:

lsmeans feeding / diff at weightbl=3.7 cl;

The DIFF option estimates the difference between the two levels of the FEEDING variable, the AT option estimates the means at a WEIGHTBL level of 3.7, and the CL option calculates the confidence limits of the means for each level of feeding method and for the difference between the two means. Continuing with the UN model the results obtained through these options of the LSMEANS statement are shown in Table 7. It can be seen that there is a difference, with respect to weight, of -0.8297 between the two levels of feeding method and in favour of the bottle-fed group.

Least Squares Means

Effect	Feeding	Weightbl	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
feeding	1	3.70	14.1753	0.2171	137	65.30	<.0001	0.05	13.7460	14.6045
feeding	2	3.70	15.0049	0.1815	137	82.65	<.0001	0.05	14.6459	15.3639

Differences of Least Squares Means

Effect	Feeding	_Feeding	Weightbl	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha
Feeding	1	2	3.70	-0.8297	0.2668	137	-3.11	0.0023	0.05

Effect	Feeding	_Feeding	Lower	Upper
Feeding	1	2	-1.3572	-0.3022

Table 7. Least squares means output for WEIGHT=3.7

4.- Conclusions

One of the main characteristics of an analysis based on the linear mixed model is the fit of the repeated measures covariance structure. This is possible as a result of new statistical programs that enable this structure to be included in the model. The choice of covariance structure can be made using graphs, by comparing covariance estimates or through the use of fit indices, the latter being the chosen option in the present study. Once the covariance structure has been defined it can be used to estimate the fixed effects.

With respect to specifying the covariance matrix by means of the PROC MIXED of the SAS system, two aspects should be highlighted. Firstly, the covariance produced by the variation between subjects is specified by means of the RANDOM statement. Secondly, when the study is based — as in the case analysed — on repeated measures the covariance structure between these measures is specified using the REPEATED statement. It is important to note that the fixed effects estimates with different covariance structures may yield the same values, even though the standard errors of these estimates can vary widely. Consequently, although the significance tests depend on the covariance structure chosen, this is not the case for the fixed effects estimates. This has a greater repercussion when the data are not balanced or the time intervals between observations are not constant.

Finally, it should be noted that additional models and analyses can be considered within the framework of linear mixed models. For example, the age effect could be analysed as a linear trend rather than as a classification variable with different values. Thus, the age variable could be treated as a continuous variable, with age effects being modelled by means of polynomial trends over time. However, as we pointed out in the Introduction, the aim of the present study was to show how to use an analytical approach that enables adequate modelling of longitudinal data using the PROC MIXED of SAS.

5.- References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control, AC-19*, 716-723.

Bock, R. D. (1989). *Multilevel analysis of educational data*. San Diego, CA: Academic Press.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models for social and behavioural research: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *HLM. Hierarchical linear and nonlinear modelling with the HLM/2L and HLM/3L programs*. Chicago, IL: Scientific Software International.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Oxford University Press.

Hoeksma, J. B., & Knol, D. L. (2001). Testing predictive developmental hypothesis. *Multivariate Behavioral Research*, *36*, 227-248.

Hox, J. J. (2002). *Multilevel analysis. Techniques and applications*. Hillsdale, NJ: Erlbaum.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*, 297-307.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics-Simulation and Computation, 27*, 591-604.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine, 19*, 1793-1819.

Luke, D. A. (2004). *Multilevel modelling*. Thousand Oaks, CA: Sage.

Plewis, I. (2001). Explanatory models for relating growth processes. *Multivariate Behavioral Research*, *36*, 207-225.

Prosser, R., Rasbash, J., & Goldstein, H. (1996). *MLn User's Guide*. London: Institute of Education.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., & Lewis, T. (2000). *A user's guide to MLwiN, Version 2.1*. London: Institute of Education.

Raudenbush, S. W. (1989). The analysis of longitudinal multilevel data. *International Journal of Educational Research, 13*, 721-740.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2000). *HLM5: Hierarchical linear and nonlinear modelling*. Lincolnwood, IL: Scientific Software International.

Sabulsky, J., Lobo, B., Agrelo, F., Berra, S., Chesta, M., Frassoni, A. M., de Ferrer, A. L., Passamonte, R., Pronsato, J., Sesa, S., & Villalba, T. (2001). *Lactancia materna y lactancia artificial. Diferencias de crecimiento en niños de la ciudad de Córdoba. Argentina*. Argentina: Triunfar S.A.

SAS Institute Inc. (2000). *SAS/STAT User's Guide, Version 8*. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2004). *SAS Online Doc 9.1.3*. Cary, NC: SAS Institute Inc.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Sheu, C. F., & Suzuki, S. (2001). Meta-analysis using linear mixed models. *Behavior Research Methods, Instruments and Computers, 33*, 102-107.

Singer, J. D. (2002). Fitting individual growth models using SAS PROC MIXED. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data: Methods and applications* (pp. 135-170). Mahwah, NJ: Lawrence Erlbaum Associates.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis. Modeling change and event occurrence*. New York: Oxford University Press.

Snijders, T. (1996). Analysis of longitudinal data using the hierarchical linear model. *Quality and Quantity, 30*, 405-426.

Verbeke, G., & Molenberghs, G. (Eds.). (1997). *Linear mixed models in practice: A SAS-oriented approach*. New York: Springer-Verlag.

Wolfinger, R. D. (1993). Covariance structure selection in general mixed models. *Communications in Statistics, Simulation and Computation, 22*, 1079-1106.

Wolfinger, R. D. (1996). Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics, 1*, 205-230.

Wolfinger, R. D. (1997). An example of using mixed models and PROC MIXED for longitudinal data. *Journal of Biopharmaceutical Statistics, 7*, 481-500.

Wright, S. P., & Wolfinger, R. D. (1997). Repeated measures analysis using MIXED models: Some simulation results. In T. G. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. G. Warren & R. D. Wolfinger (Eds.), *Modelling longitudinal and spatially correlated data* (147-158). New York: Springer-Verlag.