THE ANALYSIS OF AUXOLOGICAL DATA BY MEANS OF NONLINEAR MULTIVARIATE GROWTH CURVES
(coautore.A.M. Mineo) in corso di pubblicazione sugli atti del convegno del gruppo italiano della IFCS, Pescara luglio 1997; Springer-Verlag edit.

The Analysis of Auxological Data by Means of Nonlinear Multivariate Growth Curves

Marcello Chiodi*, Angelo M. Mineo**

* Institute of Statistics - Faculty of Economics - University of Palermo
Viale delle Scienze - 90128 - Palermo, Italy, e-mail: chiodi@unipa.it
** Department of Mechanical Technology & Production - University of Palermo
Viale delle Scienze - 90128 - Palermo, Italy, e-mail: amineo@unipa.it

Abstract: In this paper we treat the problem to analyse a data set constituted by multivariate growth curves for different subjects; thus in this context we deal with 3-way data tables. Nevertheless, it is not possible using factorial techniques proposed to deal with 3-way data matrices, because the observations are generally not equally spaced; moreover a multilevel approach founded on polynomial models is not suitable to deal with intrinsic nonlinear models. We propose a non-factorial technique to analyse auxological data sets using an intrinsic nonlinear multivariate growth model with autocorrelated errors. The application to a real data set of growing children gave easily interpretable results.

Keywords: Longitudinal studies, multivariate growth models, nonlinear regression, serial correlation, MLE, three-way data.

1. Introduction[1]

The analysis of data sets constituted by multivariate observations depending on time for different subjects is a widely studied topic; it depends on many conditions concerning the kind of data, their quality, the purpose of the analysis, and so on. In this paper, we are concerned mainly with the analysis of real data constituted by multivariate growth measures of a set of children, surveyed on different times. Therefore, at least formally, we have a 3-way data table and so we could think to use one of the specific techniques proposed to deal with 3-way data matrices, based mainly on different types of factorial decompositions. Three common methods proposed to deal with 3-way matrices "individuals x variables x occasions" are:

a) STATIS (Escoufier, 1987), that can be seen as a principal component analysis where different statistical studies with many variables are compared, by obtaining a graphical representation where the points are the studies and the proximity of the points gives a similarity among the studies;
b) the Tucker3 model (Tucker, 1966), and

[1] This research has been supported by MURST grants.

c) the PARAFAC (PARAllel FACtor) model (Harshman, 1970).

Both models b) and c) try to decompose the initial 3-way matrix by considering sets of virtual units, variables and occasions according to a minimisation function (Rizzi, Vichi, 1995). Many other methods have been proposed but few means are available to decide which method is better than the others when we have real data disposed in a 3-way matrix (Kroonenberg, 1992). Furthermore, these methods deal with 3-way matrices in which the occasions are always the same for all subjects and give generally linear decompositions. Therefore, these methods are entirely unusable for data sets constituted by individual observations with different survey times, as in our case. It is also difficult with these methods to deal properly with a serial correlation structure that could be present in the individual data.

In the following sections, we first present a data set and describe the multilevel models that could be used with growth data. Then, in section 3 we present an explicit nonlinear multivariate growth model with autocorrelated errors and in section 4 we treat the problem of the estimation of the involved parameters. In section 5 we present the results obtained by the analysis of our data set.

2. Analysis of multivariate growth curves

A longitudinal data set is constituted by k variables observed on n subjects in different occasions; in particular, our data set is a sample data set in the framework of an auxological study in order to assess growth standards: we have the weight and the height (k=2) of babies (n=64) observed in different occasions, starting mainly in the first three months from the birth and ending at an age between 3 and 5 years old. For the i-th baby and for each variable the relevant information is the observed growth curve with m_i different occasion t_{ij} ($i=1,2,...,n$; $j=1,2,...,m_i$). Lags of successive surveys are in general very different among subjects and within the same subject so that the t_{ij} are unequally spaced; also, the number of occasions m_i varies for each subject. Typical growth curves are reported in figure 1 and figure 2.

Figure 1: *Growth curves of height and weight for a single subject*

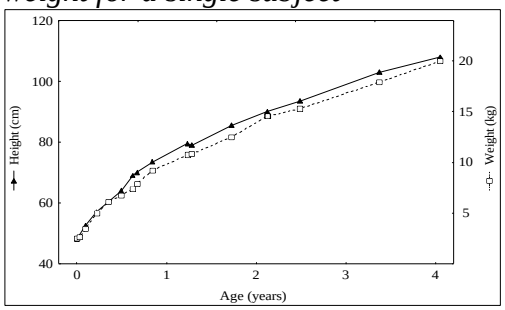
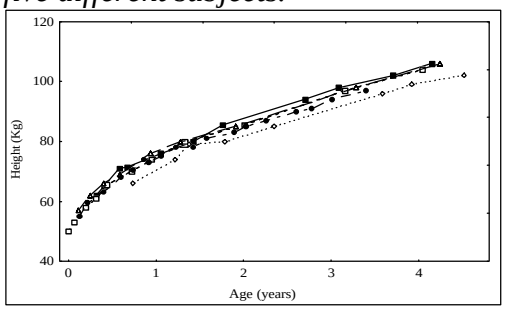


Figure 2: *Growth curves of height for five different subjects.*



Classification approaches founded on cluster analysis techniques (Mineo, 1987; Chiodi, 1989) are not useful in this context because the data cannot be seen as T

matrices of equal dimensions n×k; indeed *sections* of the 3-way data set are possible only along each subject.
Among useful approaches to the analysis of longitudinal data set, with different survey times for different subjects and with non constant time lags for each subject, the multilevel models can be taken particularly into account; in these models variability components of 1st level (different measurements of one subject) and of 2nd level (the different subjects) are considered. Besides multilevel factorial approaches (Borra, Di Ciaccio, 1996), it is interesting the 2-level growth model, proposed by Goldstein and others (1994):

$$y_{ij} = \sum_{u=0}^{p} \gamma_{iu} t_{ij}^{u} + \sum_{v=1}^{q} \alpha_v z_{ijv} + e_{ij} \qquad (1)$$

where y_{ij} is the j–th measurement on the i–th subject, γ_{iu} are the polynomial coefficients for the level 1 (the successive measurements), t_{ij} is the time of the j-th measurement on the i-th subject, z_{ijv} are the covariates, α_v are the coefficient for the covariates z's (level 2) and e_{ij} are the level 1 random terms that usually are assumed to be distributed independently with zero mean and constant variance. So these models can be considered an extension of the polynomial models for growth curves (Rao, 1965).
However, we did not use this model basically for two reasons: we have been not interested, at least in this paper, in examining random coefficient models (in the multilevel model introduced above the γ_{iu} coefficients are random at level 2 with coefficient values varying and covarying between individuals); moreover the level 1 systematic components, i.e. the time dependence of the individual measurements, have to be in our case expressly nonlinear: so we can not consider a polynomial model, even though of high degree, to obtain individual parameter estimates with a well defined biological meaning. Another opportunity is to consider a 2-level model with nonlinear systematic components on the parameters, but linear by using Taylor approximations of the first order (Milani, Bossi, 1988). In the next section, we analyse the presented data set using an explicit multivariate nonlinear growth curve model with fixed parameters.

3. Nonlinear multivariate growth model with autocorrelated errors

The main purpose of the present paper is an *exploratory* analysis of an auxological data set, to understand if the children have a similar growth with respect to the observed variables, and to understand which model can be adopted to describe the dynamic of the growth. Given very short time series, with variable time lags, we found very hard, or even impossible, to deal with this data with proper dynamic models, so that we preferred an approach based on a nonlinear growth model, which has the advantage of summarising the behaviour of each individual with a small set of parameters easily interpreted.

(coautore.A.M. Mineo) in corso di pubblicazione sugli atti del convegno del gruppo italiano della IFCS, Pescara luglio 1997; Springer-Verlag edit.

For sake of simplicity, and only to look for simple descriptive quantities which can summarise such complex data, we tried to fit, for each individual and for each variable, a general nonlinear growth curve of the family of Von Bertalanffy curves (Von Bertalanffy, 1957), that is a three parameter exponential curve:

$$E(y_t)= \gamma+ \alpha (1-e^{-\beta t}) \qquad (2)$$

Of course, the whole human growth can not be well described by only 3 parameters (Tanner, 1981): many curves have been proposed for the description of human growth even with seven parameters (Jolicoeur, Pernin, and Pontier, 1988); however, our data set concerns only the first years of human life, when growth speed decreases and this aspect is satisfactorily described by simple models. The model (2) has an easy interpretation since γ is the value of y at birth, α is a scale parameter related to the whole growth and β depends on the logarithmic growth speed: the individual fits have resulted generally better than those obtained by Gompertz or logistic curves.
The whole model is:

$$y_{ijh}=\gamma_{ih}+\alpha_{ih}[1-\exp(-\beta_{ih}t_{ij})]+\varepsilon_{ijh} \qquad i=1,2,...,n;\ j=1,2,...,m_i;\ h=1,...,k \qquad (3)$$

where y_{ijh} is the value of the h-th variable observed at the j-th occasion t_{ij} of the i-th individual, γ_{ih}, α_{ih}, β_{ih} are the parameters of the i-th individual and h-th variable, ε_{ijh} is the random error.
A peculiarity of growth curves is the possible presence of serial correlation between the measurements of an individual (Palmer, Phillips and Smith, 1991); so we assumed that random errors ε_{ijh} are normally distributed and the generic random vector $\mathbf{e}_{ih,}$ (constituted by the m_i errors of the h-th variable and the i-th individual) has covariance matrix:

$$E(\mathbf{e}_{ih}\,\mathbf{e}'_{ih})= \sigma^2_{ih}\,\mathbf{R}_{ih}, \qquad (4)$$

where σ^2_{ih} is the common variance and $\mathbf{R}_{ih}$ is a correlation matrix with generic (j,s) element $\rho_{ih_{js}}$ representing the correlation between elements of $\mathbf{e}_{ih}$ at times t_j and t_s. Of course, we need a model for the autocorrelations, in order to employ a limited number of parameters. Since the times t_{ij} are not equally spaced, we could not employ ordinary discrete time ARMA models, so that we modelled the autocorrelations according to an exponential decay (Diggle, 1988):

$$\rho_{ih_{js}} = E(\varepsilon_{ijh}\,\varepsilon_{ish})/\sigma^2_{ih} = \rho_{ih}^{|t_{ij}-t_{is}|} \qquad \rho_{ih}\geq 0. \qquad (5)$$

This is the autocorrelation function of a continuous AR(1) process (Jones, Ackerson, 1990), which allows only non negative serial correlation. At the first stage, the autocorrelations ρ_{ih} have been supposed different for each individual and each variable. Finally, we supposed that random errors $\mathbf{e}_{ih}$ are not correlated

among different individuals and different variables. Individual correlations among variables are taken into account in the systematic component of the model (3).

4. Estimation of the parameters of the model

With the assumptions of the previous section, the log-likelihood function l_{ih} for the m_i data of the i-th individual and the h-th variable is given by:

$$l_{ih}(\alpha_{ih},\gamma_{ih},\beta_{ih},\rho_{ih},\sigma^2_{ih}|\mathbf{y}_{ih})=$$
$$=-n\log(\sigma^2_{ih})/2 - \log(|\mathbf{R}_{ih}|)/2 - (\mathbf{y}_{ih}-\mathbf{f}_{ih})'\mathbf{R}_{ih}^{-1}(\mathbf{y}_{ih}-\mathbf{f}_{ih})/(2\sigma^2_{ih})$$
$$(i=1,2,...,n;\ h=1,...,k) \qquad (6)$$

being $\mathbf{y}_{ih}$ the vector of observed data and $\mathbf{f}_{ih}$ the vector of fitted data, depending on the unknown parameters α_{ih}, γ_{ih}, β_{ih}, according to the model defined by (3), and $\mathbf{R}_{ih}$ is defined through the relations (4) and (5).
In order to estimate the whole set of parameters, we have to maximise the above quantities; for sake of brevity we do not report in this paper the explicit expressions of the inverse and the determinant of $\mathbf{R}_{ih}$, since simple expressions are given by Núñez-Antón and Woodworth (1994): in fact, as in the usual case of equally spaced times and discrete time AR(1) process, the inverse of $\mathbf{R}_{ih}$ is a tridiagonal or Jacobi matrix depending only on ρ_{ih} and the set of t_{ij}, while its determinant is given by a simple factorisation.
As usual in ML estimation in regression models, we can estimate σ^2_{ih} as an explicit function of the other parameters α_{ih}, γ_{ih}, β_{ih}, ρ_{ih} and then maximise the likelihood concentrated on the latter set of parameters. In fact, the MLE s^2_{ih} of the variance component σ^2_{ih} is:

$$s^2_{ih}(\alpha_{ih},\gamma_{ih},\beta_{ih},\rho_{ih})=(\mathbf{y}_{ih}-\mathbf{f}_{ih})'\mathbf{R}_{ih}^{-1}(\mathbf{y}_{ih}-\mathbf{f}_{ih})/n, \qquad (7)$$

so that by substitution in $l_{ih}(.)$ we have the concentrated log-likelihood:

$$l_{ih}(\alpha_{ih},\gamma_{ih},\beta_{ih},\rho_{ih},s^2_{ih}(.))=-n\log((\mathbf{y}_{ih}-\mathbf{f}_{ih})'\mathbf{R}_{ih}^{-1}(\mathbf{y}_{ih}-\mathbf{f}_{ih})/n)/2-\log(|\mathbf{R}_{ih}|)/2-n/2 \qquad (8)$$

which is maximised with respect to α_{ih}, γ_{ih}, β_{ih}, ρ_{ih}, with ordinary optimisation methods. When we deal with models with reduced sets of parameters or however with constrains on the parameters, overall sample likelihood has to be used: of course the whole log-likelihood is obtained adding $l_{ih}(.)$ for all values of i and h, since we supposed the independence of the random errors among individuals and variables. Specific values of the parameters could be tested by comparing the unconstrained maximum with the maximum obtained imposing ν constraints to the parameters and then using the LR (Likelihood Ratio) test;

asymptotically -2log(LR) follows a χ^2 distribution with ν degrees of freedom, but unfortunately the number m_i of observations for each individual is generally too small in our data set, so that the χ^2 approximation to LR could be used only to give a rough judgement on the reliability of specific hypothesis.

5. Application to a real data set

The main aim of the proposed parameterisation for our data set is to deal with a 2-way data set, because the parameters of the systematic part of the model, γ_{ih}, α_{ih}, β_{ih}, summarise the third way, i.e. time. In a first stage we applied the above parameterisation to our data set, obtaining a set of 3×n×k parameter estimates: in fact we have 3 estimated parameters for each of the n=64 individuals and for each of the k=2 variables (height and weight). The analysis of the relationships between the estimates suggested some reductions in the number of parameters; for the i-th individual we put:

$\rho_{i1}=\rho_{i2}=\rho_i$ (autocorrelations are equal for the two variables but generally different among individuals);

$\beta_{i1}=\beta_{i2}=\beta_i$ (equal individual growth speeds for the two variables but generally different among individuals). A similar simplification is used by Lundbye-Christensen (1991).

Indeed the last simplification is also strongly suggested by the data, as well as the need of using all the information at disposal to estimate individual growth speeds; in fact the height has a 12% average percentage of missing data. Furthermore, the strong internal (infra-individual) linear correlation between the height and weight suggested us this simplification. The decrease of likelihood of this simplified model was not significant, so that we summarised the data set by means of the estimates of the five parameters of the systematic component: $\hat{\alpha}_{i1}$, $\hat{\gamma}_{i1}$, $\hat{\alpha}_{i2}$, $\hat{\gamma}_{i2}$ and the common slope $\hat{\beta}_i$. This estimated common individual slope $\hat{\beta}_i$ resulted to be highly correlated (R=0.95) with the individual slopes $\hat{\beta}_{i1}$ and $\hat{\beta}_{i2}$ estimated separately for the two variables.

The data did not present any evidence of difference between male and female parameters. Two individual parameter estimates appeared to be very far from the bulk of the data, so that we eliminated them from subsequent stages: they belong to children for which the above assumptions lead to unrealistic parameter estimates, since their observed growth curves are almost linear. In table 1 we report the mean and standard deviation of the individual estimates of the parameters, computed on the remaining 62 subjects.

Table 1: *Mean and standard deviation of the individual parameter estimates, computed on 62 subjects*

Estimate	$\hat{\gamma}_{i1}$	$\hat{\alpha}_{i1}$	$\hat{\gamma}_{i2}$	$\hat{\alpha}_{i2}$	$\hat{\rho}_i$	$\hat{\beta}_i$
Mean	4.48	16.69	54.58	64.78	0.05	0.42
Std. Dev.	1.19	5.48	4.99	13.97	0.06	0.20

(coautore.A.M. Mineo) in corso di pubblicazione sugli atti del convegno del gruppo italiano della IFCS, Pescara luglio 1997; Springer-Verlag edit.

An interesting aspect is the strong non-normality of the joint distribution of the estimates, as can be seen from figure 3, where we plotted the pairs of values of the estimates of α_{i1} and β_i. We see some evidence against the joint normality of the sampling distribution of the estimates, as it can be expected given the intrinsic nonlinearity of the model (Seber, Wild, 1989) also from figure 4, where we reported the likelihood contour plot of the 22nd individual with respect to same pair of parameters (α_{i1} and β_i, with i=22).

Figure 3: *Plot of 62 pairs of estimates of a_{i1} (x-axis) and b_i (y-axis)*

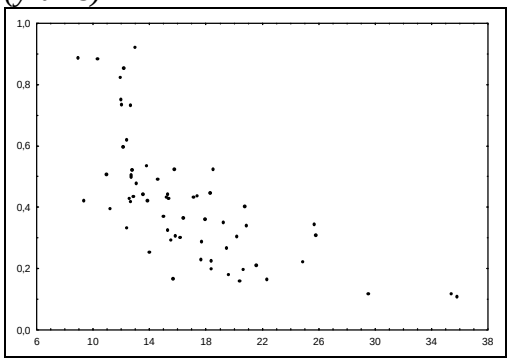
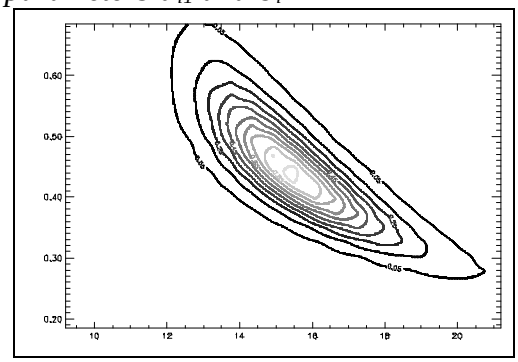


Figure 4: *Likelihood contour plot of the 22nd individual for the parameters a_{i1} and b_i*



6. Conclusion

The above analysis shows that the model (3), together with the assumptions made on serial correlations, is suitable to analyse the growth of children. The data have suggested some reduction on the number of parameters: in particular we estimated a common individual slope $\hat{\beta}_i$ and a common individual serial correlation $\hat{\rho}_i$ for both variables; even if there is still a strong collinearity among the estimates of the remaining parameters, in the present paper we do not mention any further reduction of parameters.
A strong non normality of the sampling distribution of the parameter estimates is suspected, as usual in intrinsic nonlinear models.
The obtained promising results induced us to deal, in a forthcoming paper, with random coefficient nonlinear models, in order to better deepen the study of the variability among individual growth parameters.

References

Borra S., Di Ciaccio A. (1996). Analisi fattoriale multilevel:potenzialità di analisi nell'ambito della valutazione scolastica. In *Nuove metodologie per l'analisi di dati a tre indici*; Workshop held on November, 19th, 1996, in Dipartimento di Statistica, Probabilità e Statistiche Applicate; Roma, 20-21.

Chiodi, M. (1989). The clustering of longitudinal multivariate data when time series are short. In: *Multiway data analysis*. Editors: Coppi,R. and Bolasco,S. Elsevier Science Publisher B.V. (North-Holland).

Diggle, P.J. (1988). An Approach to the Analysis of Repeated Measurements. *Biometrics*, 44, 959-971.

Escoufier, Y. (1987). Three-mode data analysis: the STATIS method, *Methods for Multidimansional Data Analysis*, ECAS, 259-272.

Goldstein H., Healy M.J.R., Rasbash J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, 13, 1643-1655.

Harshman R.A. (1970). Foundations of the PARAFAC procedure: Models and methods for an "explanatory" multi-mode factor analysis, *UCLA Working Papers in Phonetics*, 16, 1-84.

Jolicoeur, P., Pernin, M.O, Pontier, J. (1988). A Lifetime Asympotic Growth Curve for Human Height. *Biometrics*, 44, 995-1003.

Jones, R.H. and Ackerson, L.M. (1990). Serial correlation in unequally spaced longitudinal data. *Biometrika*, 77, 4, 721-731.

Kroonenberg, P.M. (1992). Three-mode component model: a survey of the literature, *Statistica Applicata*, 4, 4, 619-633.

Lundbye-Christensen, S. (1991). A Multivariate Growth Curve Model for Pregnancy. *Biometrics*, 47, 637-657.

Milani S., Bossi A. (1988). Relazione tra modelli lineari classici per lo studio dell'accrescimento somatico, *Atti della XXXIV Riunione Scientifica della Società Italiana di Statistica*, Siena 27-30 Aprile 1988, 2, 2, 77-84.

Mineo, A. (1987). Solution using clustering method. In *Data analysis:The ins and outs of solving real problems*. Editors: Janssen,J., Marcotorchino,F. and Proth,J.M.; Plenum Press, New York.

Núñez-Antón, V., Woodworth, G.G. (1994). Analysis of longitudinal data with unequally spaced observations and time-dependent correlated errors. *Biometrics*, 50, 445-456.

Palmer, M.J., Phillips, B.F., Smith, G.T. (1991). Application of Nonlinear Models with Random Coefficients to Growth Data. *Biometrics*, 47, 623-635.

Rao, C.R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, 52, 447-458.

Rizzi, A., Vichi, M. (1995). Three-way data set analysis. In: *Some Relations Between Matrices and Structures of Multidimensional Data Analysis*, Editor: Rizzi, A., Consiglio Nazionale delle Ricerche; Giardini editori, Pisa, 93-166.

Seber, G.A.F., Wild, C.J. (1989). *Nonlinear regression*. John Wiley, New York.

Tanner, J. M. (1981). *Auxologia dal feto all'uomo: la crescita fisica dal concepimento alla maturità*. Ed. italiana a cura di L.Benso, UTET, Torino.

Tucker, L.R. (1966). Some mathematical notes on three-mode factor analysis, *Psychometrika*, 31, 279-311.

Von Bertalanffy, R. (1957). Quantitative laws in metabolism and growth. *Quarterly Review of Biology*, 32, 217-231.