IDENTIFYING OUTLYING GROWTH PROFILES IN THE GROWTH OF CONIFERS

Mario Miguel Ojeda[1] and Sergio Francisco Juárez[2]
Facultad de Estadística e Informática, Universidad Veracruzana, Xalapa, México

ABSTRACT
Our objective in this paper is the detection of atypical growth profiles, which is illustrated in a growth study from 74 families of conifers. Our approach starts by fitting a 2-level linear model where we assign the measurements made on time in each family to the first level of the model, and assign the families to the second level. In order to identify atypical growth profiles we analyze the (multivariate) residuals in the second level of the fitted model. Mahalanobis distances to the origin indicate potential atypical growth profiles, however, Hadi´s more sophisticated procedure concludes that there are no outlying residuals, thus avoiding the wrong conclusion that observations with high Mahalanobis distances to the origin are necessarily outliers.

Key words: outliers in multivariate data, second level residuals in 2-level models.

MSC: 62P10.

RESUMEN
Nuestro objetivo en este artículo es la detección de perfiles de crecimiento atípicos en 74 familias de coníferas. Nuestro enfoque empieza ajustando un modelo lineal de dos niveles en el cual asignamos las mediciones hechas en el tiempo en cada familia al primer nivel del modelo y asignamos las familias al segundo nivel. Para identificar los perfiles de crecimiento atípicos, analizamos los residuos (multivariados) en el segundo nivel del modelo ajustado. Las distancias de Mahalanobis al origen indican potenciales perfiles de crecimiento atípicos. Sin embargo, el procedimiento más sofisticado propuesto por Hadi, concluye que no hay residuos atípicos, evitándose así la conclusión errónea de que distancias de Mahalanobis al origen grandes son necesariamente outliers.

1. INTRODUCTION

Multilevel linear models permit the analysis of complex nested data structures. Data structures with two levels are common in educational studies (students nested within schools), in survey data (two-stage samples), and in growth curves analysis (replications nested within individuals). These models take account of the fact that growth characteristics of individuals vary around an average trend, and that each measurement made in the individual vary around its growth trajectory. This approach permits considering two levels of variation simultaneously and also to use characteristics in both levels in order to explain the variation. Frequently, the time can be incorporated even if explanatory measurements in the level 1 are not available. One of the main advantages of these models is the possibility of considering explanatory variables associated with the second level units or group of units in order to explain the variability between the group model parameters. Moreover, the residual effects in the second level units permit to study the group distribution and to identify atypical groups. See Goldstein (1995) and Singer and Willet (2003) for references about this topic.

The rest of this article is organized as follows. In Section 2 we formulate a growth curve model for conifers using a 2-level regression model and identify outlying observations at level 1 by analyzing the residuals in the second level of the model. In Section 3 we fit the model to the data and concentrate our attention to the identification of outlying growth profiles. This leads us to the exploration and detection of outlier residuals in the second level of the model. It should be mentioned that outlier detection in multilevel model has been studied, but several problems still remain open (Langford and Lewis, 1998). The 2-level residuals are multivariate, thus we use the approach proposed by Gnanadesikan and Kettenring (1972) for the detection of multivariate outliers. Their approach consists of considering the residuals as an unstructured multivariate data

E-mail: [1]mojeda@uv.mx
[2]sejuarez@uv.mx.

set and use techniques for detection of multiple multivariate outliers. In particular, we employ the procedure proposed by Hadi (1992, 1994). We make a few remarks in Section 4.

2. THE MODEL AND RESIDUALS

We consider the following 2-level linear model: In the level 1 for each unit *i* we have

$$Y_i = X_i \beta_i + e_i, \quad i = 1,\ldots,n, \tag{1}$$

where Y_i is an m_i vector of responses, X_i is an $m_i \times p$ matrix of observable non-random predictors at level 1, β_i is a p vector of unknown level 1 coefficients and e_i is the error vector normally distributed with $E(e_i) = 0$ and $var(e_i) = \sigma^2 I_{m_i}$, where σ^2 is an unknown parameter and I_{m_i} is the identity matrix of order m_i. In the level 2 for each β_i we have

$$\beta_i = W_i \gamma + u_i \tag{2}$$

where W_i is a $p \times k$ matrix of observable non-random level-2 predictors, γ is a vector of unknown fixed parameters, and u_i is a vector of unknown normally distributed random errors with $E(u_i = 0)$ and $var(u_i) = \Omega$. Combining equations (1) and (2), we obtain the linear mixed model

$$Y_i = X_i W_i \gamma + X_i u_i + e_i,$$

with $Y_i \sim N(X_i W_i \gamma, X_i \Omega X_i^T + \sigma^2 I_{m_i})$. Finally, we assume that the random elements from different units are uncorrelated, and therefore independent, that is $cov(u_i, u_j) = 0$, $cov(e_i, e_j) = 0$ and $cov(u_i, e_j) = 0$ for $i \neq j$.

For detailed presentations of the statistical inference procedures for this model see Bryk and Raudenbush (1992), Goldstein (1995), and Searle **et al**. (1992). Some of the computer programs that implement these procedures are BMDP-5, GENMOD, HLM, ML3, and VARCL; for a review of these programs see Kreft **et al**. (1994).

To identify outlying units at level 1 we have to analyze the residuals in the second level. In order to do this, recall that we have assumed that 2-level errors follow a multivariate normal distribution, so we can consider the residuals in level 2 as a multivariate set of data, and therefore, we need procedures to identify outliers in multivariate residuals. This is not an easy problem, and no straightforward approach exists to solve it. Informal techniques for the detection of multivariate outliers are reviewed and proposed by Gnanadesikan and Kettenring (1972). In fact, we will follow their idea of considering the residuals as an unstructured multivariate sample, and then, based on such a point of view, employ multivariate techniques to explore the residuals. Let us denote as $\hat{U}$ the matrix $n \times p$ whose rows are the residual vectors $\hat{u}_i^T$. This multivariate residual matrix will be our unstructured multivariate sample. One of the Gnanadesikan and Kettering (1972) suggestions is the use of distance measures in the class of quadratic forms

$$(u_i - \bar{u})^T S^b (u_i - \bar{u}),$$

where $\bar{u}$ and S are the mean vector and the covariance matrix of $\hat{U}$, respectively. Note that, for the residuals, $\bar{u} = 0$ and $S = n^{-1}(\hat{U}^T\hat{U})^{-1}$. For b = 0, we obtain the Euclidean distance $(u_i^T u_i)^{1/2} = \|u_i\|$, and for b = -1 we have the Mahalanobis distance $(u_i^T S^{-1} u_i)^{1/2}$. These distances can be visualized in several forms, for example, a plot of the residuals projected onto their principal axis, a GH-biplot of the residuals, or an index plot of these distances, can be done. However, as Hadi (1992) points out, the Mahalanobis distance is not a robust measure due to fact that it depends on the mean vector and the covariance matrix, which are not robust. This causes two problems: “outliers do not necessarily have large values for Mahalanobis distance”, and “not all observations with large values of Mahalanobis distances are necessarily outliers”. These problems are known as masking and swamping, respectively. On the other hand, the classical methods for multivariate outlier detection (Barnett and Lewis, 1994, part III) are powerful when the data contains only one outlier observation, but, when several outliers are present, these methods do not work well anymore. This occurs because these methods loose power due to the problems of masking and swamping. Motivated for these problems, Hadi (1992) proposes a procedure for the detection of multiple outliers in multivariate data

which avoids the problems of masking and swamping. Hadi (1994) presents a modification of the procedure in Hadi (1992), which is the one we use in this work.

For reasons of space we do not present the details of Hadi´s procedure but refer the interested reader to Hadi (1992) and Hadi (1994). A simple sketch of Hadi's procedure is the following: using robust estimates of the mean vector and covariance matrix, the entire data set is divided in two subsets, called "basic subset" and "non-basic subset". The basic subset contains the observations considered as "good". Then, the basic subset is incremented adequately with the observations on the non-basic subset. The augmentation of the basic subset is done until some well defined stop criterion is satisfied. Finally, the observations in the final non basic subset are declared outliers.

3. THE APPLICATION TO GROWTH IN CONIFERS

The data set that we analyze consists of growth (height in centimeters) in 74 families of conifers with 10 replications (1 replication = 1 plant) in each family, and under greenhouse conditions. The observations were made weekly during seven weeks for some families and eight weeks for other families (between plants from the same family). Therefore, we decided to study the variability between the families using the (average) profile of growth for each family. To model each profile growth curve we identified, after and exploratory analysis, the Hoerl's function (Daniel and Wood, 1980; p. 22-23) as the best choice. Thus, for the first level we postulate the following model for each average profile growth curve:

$$\ln(y_{ij}) = \beta_{0i} + \beta_{1i}\ln t_{ij} + \beta_{2i}t_{ij} + e_{ij},\quad i = 1,\ldots,74,\ j = 1,\ldots m_i,$$

where y_{ij} is the average height of the family i at the week j and t_{ij} = j. For the second level we consider the model:

$$\beta_{0i} = \gamma_{00} + \gamma_{01}w_i + u_{0i}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11}w_i + u_{1i}$$

$$\beta_{2i} = \gamma_{20} + \gamma_{21}w_i + u_{2i}$$

where $w_i = 100(y_{i2} - y_{i1})/y_{i1}$ is an *early growth ratio percentage.* Combining equations (3) and (4) we obtain

$$\ln(y_{ij}) = \gamma_{00} + \gamma_{01}w_i + \gamma_{10}\ln t_{ij} + \gamma_{11}w_i \ln t_{ij} + \gamma_{20}t_{ij} + \gamma_{21}w_i t_{ij} + u_{01} + \ln t_{ij}u_{1i} + t_{ij}u_{2i} + e_{ij}.$$

For the random quantities we assume

$$E(e_{ij}) = 0,\ \mathrm{var}(e_{ij}) = \sigma^2,\ \mathrm{cov}(e_{ij}, e_{ik}) = 0,$$

$$E = \begin{pmatrix} u_{0i} \\ u_{1i} \\ u_{2i} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix};\quad = \mathrm{var}\begin{pmatrix} u_{0i} \\ u_{1i} \\ u_{2i} \end{pmatrix} = \begin{pmatrix} 00 & & \\ 10 & 11 & \\ 20 & 21 & 22 \end{pmatrix},$$

and $\mathrm{cov}(u_{0i}, e_{ij}) = \mathrm{cov}(u_{1i}, e_{ij}) = \mathrm{cov}(u_{2i}, e_{ij}) = 0$ for i,j = 1,…,74. The β_{ki}'s (k = 01,2), are the parameters, associated with the terms of the Hoerl's function, of the log-growth trajectory of family i. Using ML3 (see http://multilevel.ioe.ac.uk/index.html for information about ML3) we obtained the estimations in Table 1.

Table 1. Estimated parameters and their standard errors in parenthesis.

$\hat{\gamma}_{00}$	0.53	(0.0498)
$\hat{\gamma}_{10}$	- 0.694	(0.0451)
$\hat{\gamma}_{20}$	0.42	(0.0213)
$\hat{\gamma}_{01}$	0.0056	(0.032)

$\hat{\gamma}_{11}$	0.0112	(0.00289)
$\hat{\gamma}_{21}$	- 0.000318	(0.00123)

The estimated Variance-Covariance components are $\hat{\sigma}^2 = 0.000406$ and

$$\hat{\Omega} = \begin{bmatrix} 0.0232(0.00439) & & \\ 0.0121(0.00305) & 0.00201(0.00386) & \\ -0.00527(0.00141) & -0.0037(0.00156) & 0.00279(0.000737) \end{bmatrix},$$

We also obtained the second level residual $\hat{u}_i^T = (\hat{u}_{0i}, \hat{u}_{1i}, \hat{u}_{2i})$, (i = 1,…,74) which we arranged in a 74×3 multivariate residual matrix. Figure 1 shows the principal components scatter-plot of the residuals. With this plot, we can say that those points lying outside the confidence ellipse are potential outliers. Figure 2 is the index plot of the Mahalanobis distances $(\hat{u}_i^T S^{-1} \hat{u}_i)^{1/2}$. At the 0.05 significance level, the critical value is $(\chi^2_{3,0.975})^{1/2} = 3.06$. So the Mahalanobis distance identifies some observations as potential outliers. However, a more elaborated analysis using the method of Hadi concludes that there are no outlier residuals (Hadi's procedure is implemented in STATA, STATA Reference Manual, vol. two, p. 432-437. We have to remark that STATA is the only software that implements Hadi's procedure). This fact indicates the likely existence of swamping in the residual matrix, in other words, the presence of non outlying residuals with high Mahalanobis distances to the origin.

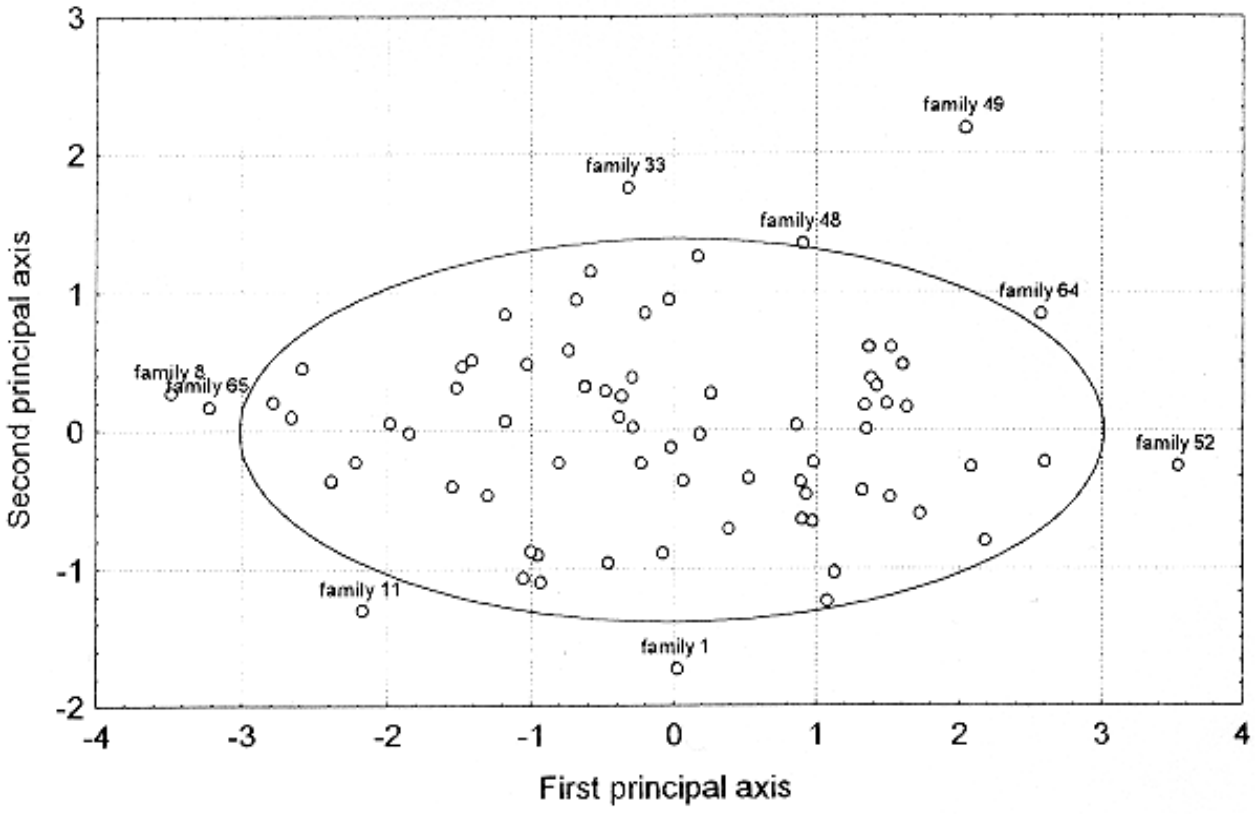

Figure 1. First two principal components scatter-plot with a 95% confidence ellipse of the level 2 residual matrix obtained from the fit model. The proportion of explained variance is 0.97.

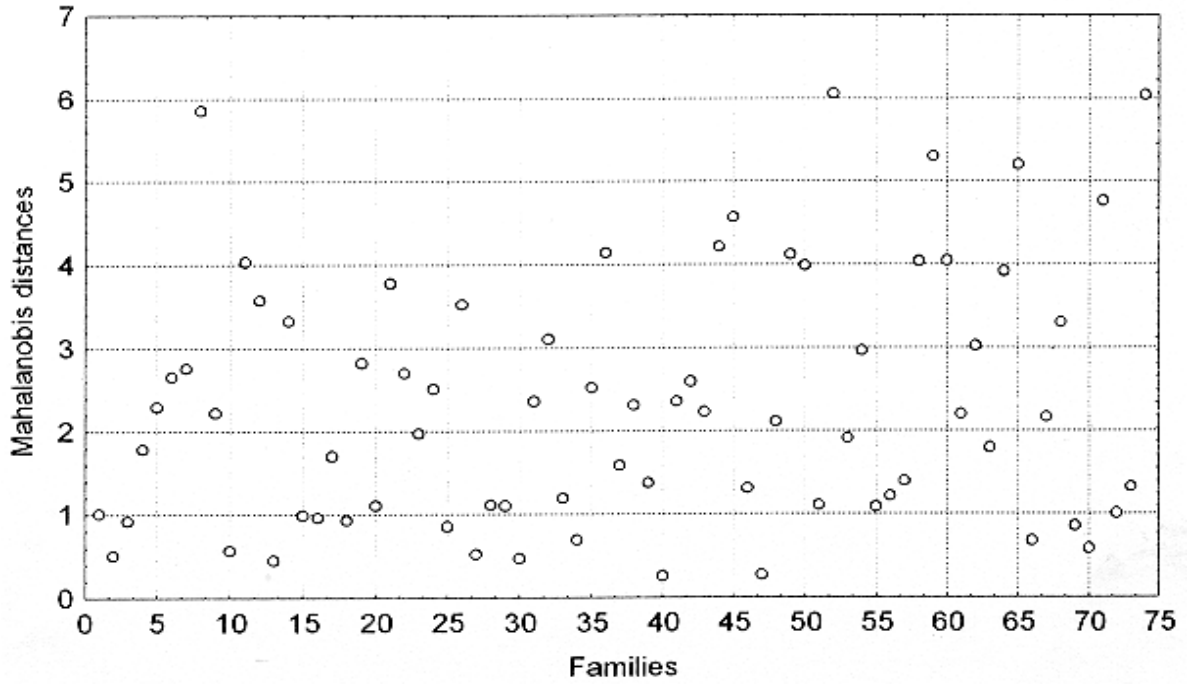

Figure 2. Index plot of the Mahalanobis distances for the level 2 residuals of the fit model.

4. CONCLUSION

Multilevel linear models permit to approach very realistically the modeling process in several repeated measures situations. In growth curve analysis the type of models are for concrete applications (see Hand and Crowder, 1996; Vonesh, and Chinchilli, 1997), but the data analysis process in this context requires diagnostic tools for evaluating the atypical individuals, which could be a difficult problem (Shi and Ojeda, 2004). Exploratory and descriptive analysis of residuals in a 2-level regression model fit help in a preliminary identification of candidate outliers, but a formal testing of real outlying effect is necessary. Hadi´s test is a powerful technique for the identification of real multiple multivariate outliers, as we illustrated in the application.

REFERENCES

BARNETT, V. and LEWIS, T. (1994): **Outliers in Statistical Data**, 3rd edition. Wiley, New York. 584 p.

BRYK, A.S. and D.J. RAUDENBUSH (1992): "Hierarchical Linear Models: Applications and Data Analysis Methods", **Sage Publications**, Newbury Park. 265 p.

DANIEL, C. and F.S. WOOD (1980): **Fitting Equations to Data**, 2nd edition. Wiley, New York. 480 p.

GNANADESIKAN, R. and J.R. KETTENRING (1972): "Robust estimates, residuals, and outlier detection with multiresponse data", **Biometrics** 28, 81-124.

GOLDSTEIN, H. (1995): **Multilevel Statistical Models**, 2nd edition. Edward Arnold, London. 178 p.

HADI, A.S. (1992): "Identifying multiple outliers in multivariate data", **J.R.S.S**. B, 54, 761-771.

_____ (1994): "A modification of a method for the detection of outliers in multivariate samples", **J.R.S.S**. B, 56, 393-396.

HAND, D. and M. CROWDER (1996): **Practical Longitudinal Data Analysis**. Chapman and Hall, London. 239 p.

KREFT, I.G.G.; J. de LEEUW and R. VAN DEER LEEDEN (1994): "Review of five Multilevel Analysis Programs: BMDP-5V, GENMOD, HLM, ML3, VARCL", **The American Statistician** 48(4), 324-335.

LANGFORD, I.H. and T. LEWIS (1998): “Outliers in multilevel data”, **J.R.S.S.** A, 161, 121-160.

SEARLE, S.R.; G. CASELLA and C.E. MCCULLOCH (1992); **Variance Components**. Wiley, New York. 501 p.

SHI, L. and M.M. OJEDA (2004): “Local Influence in Multilevel Regression for Growth Curves”, **Journal of Multivariate Analysis**. (In press).

SINGER, J. D. and J.B. WILLET (2003): **Applied Longitudinal Data Analysis**. Oxford University Press, New York. 644 p.

STATA Reference Manual, Release 4. Volume Two. Stata Press. College Station, Texas. 523 p.

VONESH, E.F. and V.M. CHINCHILLI (1997): **Linear and Nonlinear Models for the Analysis of Repeated Measurements**. Marcel Dekker, New York. 560 p.