A TWO-STAGE ESTIMATOR OF INDIVIDUAL REGRESSION COEFFICIENTS IN MULTIVARIATE LINEAR GROWTH CURVE MODELS

por

Gabriela Beganu[1]

Resumen

Beganu, G: A two-stage estimator of individual regression coefficients in multivariate linear growth curve models. Rev. Acad. Colomb. Cienc. **30** (117): 549-554, 2006. ISSN 0370-3908.

Se considera una familia de modelos lineales de curvas de crecimiento multivariadas con efectos aleatorios. El propósito del artículo es describir un método de cálculo para la estimación de coeficientes de regresión individuales cuando las componentes de covarianza sean conocidas o no. En el segundo caso, la matriz de covarianza de los datos se estimará usando una versión generalizada del método III de Henderson mediante proyecciones ortogonales sobre subespacios lineales que corresponden al modelo y se obtiene un estimador en dos etapas para los coeficientes individuales de regresión. Se presenta la estimación de efectos fijos y aleatorios usando un abordaje bayesiano.

Palabras clave: Estimador empírico de Bayes estimado, estimador cuadrático insesgado, proyección ortogonal.

Abstract

A family of multivariate linear growth curve models with random effects is considered. The purpose of this article is to describe a computational method required for estimation of individual

1 Department of Mathematics, Academy of Economic Studies, Str. Piaþa Romanã Nr. 6, Bucharest, Romania, e-mail: Gabriela_beganu@yahoo.com. 2000 Mathematics Subject Classification: Primary 62H12, 62J12. Secondary 62H40.

regression coefficients when the covariance components are known or unknown. In the second case the covariance matrix of the data will be estimated by a generalized version of the Henderson method III using the orthogonal projections onto linear subspaces corresponding to the model and a two-stage esstimator of individual regression coefficients is obtained. The estimation of fixed and random effects is presented in a Bayesian approach.

Key words: Estimated empirical Bayes estimator, quadratic unbiased estimator, orthogonal projection.

1. Introduction

This article is concerned with estimation of regression coefficients in a multivariate linear growth curve model with a multivariate random effect when the covariance structure of the model is known or unknown. Since the model also contains fixed effects, it is considered to be a mixed linear model. The main emphasis of the proposed analysis is on estimation of the regression coefficients in a Bayesian approach by means of a two-stage estimator constructed with quadratic unbiased estimators of the covariance components ([5]). A family of multivariate linear growth curve models with random effects is considered as a generalized linear model of **Potthoff** and **Roy** [22]. **Reinsel** [24] assumed that, for each of n individual sampling units having m characteristics measured at each of p occasions, the $pm \times 1$ observable random vector is

$$y_k = (X \otimes I_m)Ba_k + (X \otimes I_m)\lambda_k + e_k, \quad k = 1, \ldots n \quad (1)$$

The within-individual and between-individuals design matrices X and $A' = (a_1, \ldots a_n)$ are known $p \times q$ and $r \times n$ matrices of full column rank $q \leq p$ and $r < n$, respectively and B is a $qm \times r$ matrix of unknown regression parameters. The assumption made by **Reinsel** [24], [25] regarding the covariance structure of the random effect can be changed without loss of generality as follows: λ_k is a $qm \times 1$ vector obtained from an $m \times q$ random matrix having columns identically distributed as $N(0, \Sigma_\lambda)$. Then the random vectors λ_k are independent of each other and of the error vectors e_k, which are distributed as $N(0, I_p \otimes \Sigma_e)$, $k = 1 \ldots n$. I_m is the $m \times m$ identity matrix and $\otimes$ denotes the Kronecker matrix product $(C \otimes D = (c_{i,j}D))$. Then the independent random vectors y_k have the normal distribution with the expected value

$$E(y_k) = (X \otimes I_m)Ba_k \quad (2)$$

and the covariance matrix

$$\operatorname{cov}(y_k) = V = (XX') \otimes \Sigma_\lambda + I_p \otimes \Sigma_e, \quad k = 1, \ldots n \quad (3)$$

where Σ_λ and Σ_e are $m \times m$ positive definite matrices. The individual regression coefficients

$$\beta_k = Ba_k + \lambda_k, \quad k = 1 \ldots n \quad (4)$$

may serve to characterize some aspects of an individual's growth [10] or in prediction of future observations for a given individual [25]. The problem of estimation of individual regression coefficients is of particular interest, β_k being composed of both fixed and random effects. Hence the estimation for model (1) can be based either on ordinary least squares and maximum likelihood methods, or on empirical Bayes methodology [20]. The classical approach uses the maximum likelihood estimation of the fixed effects B and of the covariance components Σ_λ and Σ_e from the marginal normal distribution of $Y' = (y_1, \ldots y_n)$ having the mean $(X \otimes I_m)BA'$ and the covariance matrix of vec Y' given by $I_n \otimes V$, where V is the covariance matrix in (3) and vec $Y' = (y_1', \ldots y_n')'$. The realized values of the random effects $\Lambda' = (\lambda_1, \ldots \lambda_n)$ can be estimated using the generalized version of the Gauss-Markov theorem [13], [14]. A possible alternative [13] is based on a Bayesian estimation of the model parameters (4) and it was used in [26], [23], [7], [27]. The empirical Bayes estimators of B and $\lambda_k, k = 1, \ldots n$, will be the estimated means of the posterior distribution of Y'. When the covariance components of the observation vectors are unknown, the corresponding estimators could be found in [10] and [18] by the EM algorithm (see [9]). In [8] and [19] the ordinary least squares residuals were used for computing the maximum likelihood estimators (MLE) and the restricted maximum likelihood estimators (REMLE) of the variance and covariance components in linear models for serial measurements that include both growth curves and repeated

measures models. A generalized version ([5], [6]) of the Henderson method III (also called the method of fitting constants [15]) was chosen to obtain the quadratic unbiased estimators of Σ_λ and Σ_e and it will be presented in Section 2 for the specific case of model (1). In order to obtain the estimators of regression coefficients when the covariance matrix of data is unknown, a multistage procedure is required. An estimated generalized least squares estimator (GLSE) of regression coefficients in multivariate mixed linear models is calculated in three steps in [8]. Two-stage procedures are used in [16], [17] and [23] to estimate the fixed effects and the realized values of the random effects in some special cases of general mixed linear models.

A two-stage procedure is employed to estimate the individual regression coefficients (4), when the covariance components are unknown and the corresponding estimators are obtained by the generalized Henderson method III. Section 3 deals with determining of two-stage estimators of regression coefficients (4), when the multivariate growth curve model (1) is balanced and unbalanced (missing data), respectively.

2. Estimation of the covariance components

Various methods of estimation of the covariance components are available for mixed linear models, but I confine myself to the Henderson method III. Unlike ML and REML estimations, the Henderson method III is a computationally simple method and does not require the assumption of normality. The development of this method can be made if the model (1) is put into an appropriate form using a coordinate-free approach ([2], [3], [4]). A matriceal form of model (1) is

$$Y = AB'(X' \otimes I_m) + \Lambda(X' \otimes I_m) + E \qquad (5)$$

where $\Lambda' = (\lambda_1, \dots \lambda_n)$ and $E' = (e_1, \dots e_n)$ are $qm \times n$ and $pm \times n$ random matrices, respectively. From the relations (2) and (3) we obtain that

$$E(Y) = AB'(X' \otimes I_m)$$

$$\operatorname{cov}(\operatorname{vec} Y) = V \otimes I_n$$

Some notations are required in the sequel [12]. Let $\mathcal{L}_{n,m}$ be the real vector space of linear transformations on R^n to R^m endowed with the inner product $(C, D) = \operatorname{tr}(CD')$ for every $C, D \in \mathcal{L}_{n,m}$ and let $R(C) \subset R^m$ be the linear space spanned by the columns of $C \in \mathcal{L}_{n,m}$. If $A \in \mathcal{L}_{r,n}$ and $X \in \mathcal{L}_{q,p}$ are known design matrices of model (5), then let $\Omega = \{AB'(X' \otimes I_m) \mid B \in \mathcal{L}_{r,qm}\}$ be a linear manifold of $\mathcal{L}_{pm,n}$ such that $E(Y) \in \Omega$ for $B \in \mathcal{L}_{r,qm}$. In a similar way the linear manifold Θ can be considered such that $E(Y) = AB'(X' \otimes I_m) + \Lambda(X' \otimes I_m) \in \Theta$ for every $B \in \mathcal{L}_{r,qm}$ and $\Lambda \in \mathcal{L}_{qm,n}$, when the random effects are considered to be fixed effects. Hence $P_A = A(A'A)^{-1}A'$ and $P_X = [X(X'X)^{-1}X'] \otimes I_m$ are the othogonal projections onto $R(A)$ and $R(X \otimes I_m)$, respectively, it is easy to find that $P_1 = P_A \otimes P_X$ (see [11]) and $P_2 = I_n \otimes P_X$ (see [6]) are the orthogonal projections onto Ω and Θ, respectively. Then

$$I_{npm} - P_2 = I_n \otimes M_X \qquad (6)$$

and

$$P_2 - P_1 = M_A \otimes P_X \qquad (7)$$

where $M_\Lambda = I_n - P_\Lambda$ and $M_X = I_{pm} - P_X$ are the orthogonal projections onto the orthogonal complements of $R(A)$ and $R(X \otimes I_m)$, respectively. The quadratic forms used in the generalized Henderson method III ([5]) corresponding to the model (5) will have the symmetric matrices (6) and (7) and will be obtained from the following

$$\begin{aligned}&[(I_{npm} - P_2)Y]'[(I_{npm} - P_2)Y] \\ &\quad = [(I_n \otimes M_X)Y]'[(I_n \otimes M_X)Y] \\ &\quad = (YM_X)'(YM_X) = M_X Y'Y M_X\end{aligned} \qquad (8)$$

and

$$\begin{aligned}&[(P_2 - P_1)Y]'[(P_2 - P_1)Y] \\ &\quad = [(M_A \otimes P_X)Y]'[(M_A \otimes P_X)Y] \\ &\quad = (M_A Y P_X)'(M_A Y P_X) = P_X Y' M_A Y P_X\end{aligned} \qquad (9)$$

where there were used the definition of the Kronecker operators product $(P \otimes Q)C = PCQ'$ (see [11]) and the properties of the orthogonal projections) P_A and P_X to be symmetric and idempotent operators. It is also proved ([6]) that (6) and (7) are the matrices of the quadratic forms corresponding to model (5) founded by an iterative method of estimation based on Gram-Schmidt orthogonalization process of design matrices in mixed linear models. The expected values of the random quadratic forms (8) and (9) can be found from [21] and using [1] as

$$\begin{aligned}&E(M_X Y'Y M_X) \\ &= M_X[(X \otimes I_m)BA'AB'(X' \otimes I_m) + n \cdot V]M_X \\ &= n \cdot M_X V M_X = n \cdot M_X(I_p \otimes \Sigma_e)M_X\end{aligned} \qquad (10)$$

and

$$E(P_X Y' M_A Y P_X) = P_X[(X \otimes I_m)BA'M_A AB'(X' \otimes I_m) + \text{tr } M_A \cdot V]P_X = (n-r)\cdot P_X V P_X = (n-r)(X \otimes I_m)[I_q \otimes \Sigma_\lambda + (X'X)^{-1} \otimes \Sigma_e](X' \otimes I_m) \quad (11)$$

The Henderson method III consists in equating the quadratic forms (8) and (9) to their expected values (10) and (11), respectively. It was proved in [5] that the equation obtained by this method are consistent. Then the solution $\hat{\Sigma}_\lambda$ and $\hat{\Sigma}_e$ of the estimating equations

$$\begin{cases} Y'Y = n\cdot(I_p \otimes \Sigma_e) \\ \left[(X'X)^{-1}X' \otimes I_m\right] Y' M_A Y \left[X(X'X)^{-1} \otimes I_m\right] \\ \qquad = (n-r)[I_q \otimes \Sigma_\lambda + (X'X)^{-1} \otimes \Sigma_\epsilon] \end{cases} \quad (12)$$

are the quadratic unbiased estimators of covariance components. The difference between the solution of (12) and the estimators obtained in [24], [25] by ANOVA method also comes from the assumption regarding the covariance structure of the random effect.

3. Estimation of the individual regression coefficients

a) Balanced case. Using the variability between and within sampling units, the estimators based on the data from a single unit can be improved by an appropriate use of data from the remaining units. Thus empirical Bayes methodology will be used for finding the estimators of fixed and random effects when the covariance matrix of data is known or unknown by means of the posterior expected value and posterior covariance. When the covariance components Σ_λ and Σ_e are known, then the fixed effects B and the realized values of the random vectors $\lambda_k, k = 1,\dots n$, are the unknown parameters of model (1). The MLE of B, which is also the ordinary least squares estimator of B,

$$\hat{B} = [(X'X)^{-1}X' \otimes I_m]Y'A(A'A)^{-1} \quad (13)$$

was given in [24] and [19]. Thus $\hat{B}$ is the best linear unbiased estimator of B because maximizes the likelihood based on the marginal distribution of y_k, $k = 1,\dots n$. The estimator of the realized value of λ_k is derived by the generalized Gauss-Markov theorem [13] as

$$\hat{\lambda}_k = (X' \otimes \Sigma_\lambda)V^{-1}[y_k - (X \otimes I_m)\hat{B}a_k] \quad (14)$$

which is also the empirical Bayes estimator of λ_k expressed by

$$\hat{\lambda}_k = E(\lambda_k \mid y_k, V, \hat{B}) \quad (15)$$

that is the conditional mean of λ_k given y_k, the known covariance V of y_k and the unknown parameters B estimated by (13) , $k = 1,\dots n$. The expression (14) of $\hat{\lambda}_k$ can be found from (15) using formulas given in [28]. Thus the estimator of the individual regression coefficients β_k is

$$\hat{\beta}_k = \hat{B}a_k + \hat{\lambda}_k, \quad k = 1,\dots n \quad (16)$$

with $\hat{B}$ and $\hat{\lambda}_k$ given by (13) and (14), respectively. When the covariance matrix V of data is unknown, a two-stage estimator of β_k will be derived by the following computational steps: - the covariance components due to the random effect and the unobservable errors are estimated using the generalized version of the Henderson method III. Then $\hat{\Sigma}_\lambda$ and $\hat{\Sigma}_e$ are the solution of (12); - the GLSE of the fixed effects B in model (1) will be

$$\hat{B}(\hat{V}) = [(X' \otimes I_m)\hat{V}^{-1}(X \otimes I_m)]^{-1}(X' \otimes I_m)\hat{V}^{-1}Y'A(A'A)^{-1} \quad (17)$$

where $\hat{V} = (XX') \otimes \hat{\Sigma}_\lambda + I_p \otimes \hat{\Sigma}_e$. The estimated empirical Bayes estimator of the realized values of random effect λ_k will be the following conditional mean

$$\hat{\lambda}_k(\hat{V}) = E(\lambda_k \mid y_k, \hat{V}, \hat{B}(\hat{V})) = (X' \otimes \hat{\Sigma}_\lambda)\hat{V}^{-1}[y_k - (X \otimes I_m)\hat{B}(\hat{V})a_k], \quad (18)$$

$k = 1,\dots n$. The two-stage estimator of individual regression coefficients (4) of model (1) will be

$$\hat{\beta}_k(\hat{V}) = \hat{B}(\hat{V})a_k + \hat{\lambda}_k(\hat{V}), \quad k = 1,\dots n \quad (19)$$

Laird and **Ware** [18] showed that the use of a combination of an estimated empirical Bayes estimator (18) corresponding to the random effects is agreed in the literature for some choise of $\hat{V}$, once $\hat{V}$ is available.

b) Missing data case of model (5). The importance of studying the missing data models should be clear to any researcher carrying out a large scale multivariate experiments. Such experiments do not always yield complete data (i.e. with no missing observations). Thus the researcher is frequently faced with the necessity of analyzing the "incomplete" data. A generalisation of model (5) allows for missing data in the sense that different individuals may not have observations at the same time points. In this case it is supposed that the n individuals with m characteristics are divided into c disjoint sets $S_1,\dots S_c$ with $n_1,\dots n_c$ experimental

units, respectively. On each individual in S_i measurements are taken on $p_i \leq p$ response variates and two different sets S_i and $S_{i'}$ can not have measurements at the same time, although they have $p_i = p_{i'}$. Let Y_i be the $n_i \times p_i m$ random matrix of observations in the set S_i, $i = 1, \ldots c$. Analogous to model (5), we assume that the growth curve model corresponding to S_i is

$$Y_i = A_i B'[(X'C_i') \otimes I_m] + \lambda_i[(X'C_i') \otimes I_m] + E_i \quad (20)$$

where A_i is an $n_i \times r$ between-individuals design matrix, C_i is an $p_i \times p$ incidence matrix of 0's and 1's and λ_i is an $n_i \times qm$ matrix of random effects, $i = 1, \ldots c$. The rows of Y_i are independent and normally distributed as

$$E(Y_i) = A_i B'[(X'C_i') \otimes I_m]$$
$$\operatorname{cov}(\operatorname{vec} Y_i) = V_i \otimes I_{n_i}$$
$$= [(C_i X X' C_i') \otimes \Sigma_{\lambda_i} + I_{p_i} \otimes \Sigma_{e_i}] \otimes I_{n_i}$$

and Y_i and $Y_{i'}$ are independent for $i \neq i'$. The empirical Bayes estimators of individual regression coefficients in the multivariate growth curve model with missing data can be obtained for every set S_i of observations using the formulae (16) and (19) corresponding to model (20), or rewritting the model by means of the vec operator as

$$\operatorname{vec} Y = D \operatorname{vec} B + G \operatorname{vec} \Lambda + \operatorname{vec} E$$

where

$$D = \begin{pmatrix} (C_1 X) \otimes I_m \otimes A_1 \\ \vdots \\ (C_c X) \otimes I_m \otimes A_c \end{pmatrix},$$

$$G = \begin{pmatrix} (C_1 X) \otimes I_m \otimes I_{n_1} \\ \vdots \\ (C_c X) \otimes I_m \otimes I_{n_c} \end{pmatrix}$$

and vec B, vec Λ and vec E are obtained like vec $Y = ((\operatorname{vec} Y_1)', \ldots (\operatorname{vec} Y_c)')'$. Then, assuming the normality of the observations, the expected mean and the positive definite covariance matrix are

$$E(\operatorname{vec} Y) = D \operatorname{vec} B$$

$$\operatorname{cov}(\operatorname{vec} Y) = \operatorname{diag}(V_i \otimes I_{n_i})$$

Therefore the multivariate growth curve model with missing data is of the form of the univariate linear regression model. A two stage estimator of individual regression coefficients is recommended to be used in mixed linear models because it uses all the information contained in data and takes into account the random nature of the effects.

References

[1] **Baksalary, J.K. and Kala, R.** (1979), *Criteria for estimability in multivariate linear models*, Math. Operationsforsch. Statist. ser. Statist., **7**, 5–9.

[2] **Beganu, G.** (1987), *Estimation of regression parameters in a covariance linear model*, Stud. Cerc. Mat., **39**, 3–10.

[3] **Beganu, G.** (1987), *Estimation of covariance components in linear models. A coordinate-free approach*, Stud. Cerc. Mat., **39**, 228–233.

[4] **Beganu, G.** (2003), *The existence conditions of the best linear unbiased estimators of the fixed factor effects*, Econom. Cumput. Econom. Cybernet. Studies and Research, **36**, 95–102.

[5] **Beganu, G.** *Quadratic estimator of covariance components in a multivariate mixed linear model*, Statistical Methods and Applications (to appear).

[6] **Beganu, G.** (2005), *On Gram-Schmidt orthogonalizing process of design matrices in linear models as estimating procedure of covariance components*, Rev. R. Acad. Cien. Serie A. Mat., **99**, 187–194.

[7] **Chen, Z. and Dunson, D. B.** (2003), *Random effects selection in linear mixed models*, Biometrics, **59**, 762–771.

[8] **De Gruttola, V.; Ware, J. H. and Louis, T.A.** (1987), *Influence analysis of generalized least squares estimators*, J. Amer. Statist. Assoc., **82**, 911–917.

[9] **Dempster, A. P.; Laird, N. M. and Rubin, D. B.** (1977), *Maximum likelihood from incomplete data via the EM algorithm*, J.Royal Statist. Soc., Ser. B, **39**, 1–38.

[10] **Dempster A. P.; Rubin, D. B. and Tsutakawa, R. K.** (1981), *Estimation in covariance components models*, J.Amer. Statist. Assoc., **76**, 341–353.

[11] **Eaton, M. L.** (1970), *Gauss-Markov estimation for multivariate linear models: A coordinate-free approach*, Ann. Math. Statist., **41**, 528–538.

[12] **Halmos, P. R.** (1958), *Finite Dimensional Vector Spaces*, 2nd ed., Van Nostrand, Princeton, New Jersey.

[13] **Harville, D. A.** (1976), *Extension of the Gauss-Markov theorem to include the estimation of random effects*, Ann. Statist., **4**, 384–395.

[14] **Harville, D. A.** (1977), *Maximum likelihood approaches to variance component estimation and to related problems*, J. Amer. Statist. Assoc., **72**, 320–340.

[15] **Henderson, C. R.** (1953), *Estimation of variance and covariance components*, Biometrics, **9**, 226–252.

[16] **Khan, S.** and Powell J.L. (2001), *Two-step estimation of semiparametric censored regression models*, J. Econometrics, **103**, 73–110.

[17] **Khuri, A. I.** (1992), *Response surface models with random block effects*, Technometrics, **34**, 26–37.

[18] **Laird, N. M. and Ware, J. H.** (1982), *Random-effects models for longitudinal data*, Biometrics, **38**, 963–974.

[19] **Lange, N. and Laird, N. M.** (1989), *The effect of covariance structure on variance estimation in balanced growth-curve models with random parameters*, J.Amer. Statist. Assoc., **84**, 241–247.

[20] **Morris, C.N.** (1983), *Parametric empirical Bayes inference: Theory and applications*, J.Amer. Statist. Assoc., **78**, 47–55.

[21] **Neudecker, N.** (1990), *The variance matrix of a matrix quadratic form under normality assumptions. A derivation*

based on its moment-generating function, Math. Operationsforsch. Statist., ser. Statistics, **3**, 455–459.

[22] **Potthoff, R. F. and Roy, S. N.** (1964), *A generalized multivariate analysis of variance model useful especially for growth curve problems*, Biometrika, **51**, 313–326.

[23] **Prassad, N. G. N. and Rao, J. N. K.** (1990), *The estimation of the mean squared error of small-area estimators*, J.Amer. Statist. Assoc., **85**, 163–171.

[24] **Reinsel, G.** (1982), *Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure*, J.Amer. Statist. Assoc., **77**, 190–195.

[25] **Reinsel, G.** (1984), *Estimation and prediction in a multivariate random effects generalized linear model*, J.Amer. Statist. Assoc., **79**, 406–414.

[26] **Reinsel, G.** (1985), *Mean squared error properties of empirical Bayes estimators in a multivariate random effects general linear model*, J.Amer. Statist. Assoc., **80**, 642–650.

[27] **Sala-i-Martin, X., Doppelhofer, G. and Miller, R. I.** (2004), *Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach*, American Economic Review, **94**, 813 835.

[28] **Smouse E. P.** (1984), *A note on Bayesian least squares inference for finite population models*, J.Amer. Statist. Assoc., **79**, 390–392.

Recibido el 13 de marzo de 2006

Aceptado para su publicación el 19 de agosto de 2006