On modeling HIV data using bivariate linear mixed model

E. Eteng [*1], E. O. Effanga[1] and M. E. Nja[2]

ABSTRACT

Bivariate linear mixed models are useful when analyzing longitudinal data of two associated markers. In this paper, we present a bivariate linear mixed model including random effects or first-order auto-regressive process and independent measurement error for both markers. We fitted these models using SAS Proc MIXED.

INTRODUCTION

Longitudinal data are often collected in epidemiological studies, especially to study the evolution of biomedical markers. Thus, linear mixed models Laird Ware (1982). When several markers are measured repeatedly, longitudinal multivariate models could be used like in econometrics. However, this extension of univariate models is rarely used in biomedicine although it could be useful to study the joint evolution of biomarkers . For instance, in HIV infection, several markers are available to measure the quantity of virus (plasma viral load noted HIV RNA), the status of immune system (CD4+ T lymphocytes which are a specific target of the virus CD8-T lymphocytes) Or the inflammation process (β_2 microglobuline).

These markers are associated as the infection measure by HIV RNA induces inflammation and the destruction of immune cells. Several authors have developed methods to fit evolution of CD4 and CD8 cells Shah et al (1997) or CD4 and β_2 microglobuline Sy et al (1997) used the Fisher scoring method to fit a bivariate linear random effects model including on Integrated Orstein-Uhlenbeck process (IOU) . IOU is a stochastic process that includes Brownian motion as special limiting case.

In this paper, we propose some tricks to use SAS MIXED procedure in order to fit multivariate linear mixed models to multivariate longitudinal Gaussian data. SAS MIXED procedure uses Newton-Raphson algorithm known to be faster than the EM algorithm Lindstrom et al (1988).

In sections 2 and 3, we present bivariate linear mixed models used in SAS to fit these models. In section 4, we apply these models to study the joint evolution of HIV RNA and CD4+ T lymphocytes in a cohort of HIV-1 infected patients (APROCO) treated with highly active antiretroviral treatment.

MODEL FOR BIVARIATE LONGITUDINAL GAUSSIAN DATA

Let $Y_i = \begin{bmatrix} Y_i^1 \\ Y_i^2 \end{bmatrix}$, the response vector for

The i, with Y_i^k the $n_i^k - vector$ of measurements of the marker $k(k=1,2)$ with $n_i^1 = n_1^2 = n_i$. If the two markers are independent, we can use the two following models

$$\begin{cases} Y_i^1 = X_i^1\beta^1 + Z_i^1\gamma_i^1 + W_i^1 + \varepsilon_i^1 \\ Y_i^2 = X_i^2\beta^2 + Z_i^2\gamma_i^2 + W_i^2 + \varepsilon_i^2 \end{cases} \quad (1)$$

$\varepsilon_i^1 \sim N(0, \sigma_{\varepsilon^1}^2 I_n), \gamma_i^1 \sim N(0, G^2), W_i^1 \sim N(0, R_i^1)$ and

$\varepsilon_i^2 \sim N(0, \sigma_{\varepsilon^2}^2 I_n), \gamma_i^2 \sim N(0, G^2), W_i^2 \sim N(0, R_i^2)$ (2)

where X_i^k is a $n_i \times p^k$ design matrix which is usually a subset of X_i^k, γ_i^k is a q^k - vector of individual random effects with $q^k \le p^k \bullet .W_i^k$ is a vector of realization of a first order auto-regressive process $w_i^k(t)$ with covariance given by

*Corresponding author
Manuscript received by the Editor January 14, 2008; revised manuscript accepted November 12, 2008.
[1]Department of Mathematics /Statistics & Computer Science, University of Calabar, Calabar, Nigeria
[2]Department of Mathematic /Statistic Cross River State University of Technology, Calabar

Data modeling using bivariate linear

$R_i^k(s,t) = \sigma^2_{w^k} e^{\lambda|t-s|}$ and I_n is a $n_i \times n_i$ identity matrix. To take into account correlation between both markers, one could use the following bivariate linear mixed model

$$Y_i = X_i\beta + Z_i\gamma_i + W_i + \varepsilon_i \text{ with } \begin{cases} \varepsilon_i \sim N(0,\Sigma_i) \\ W_i \sim N(0,R_i) \\ \gamma_i \sim N(0,G) \end{cases} \quad (3)$$

where $X_i = \begin{bmatrix} X_i^1 & 0 \\ 0 & X_i^2 \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, Z_i = \begin{bmatrix} Z_i^1 & 0 \\ 0 & Z_i^2 \end{bmatrix}$

$\gamma_i^1 = \begin{bmatrix} \gamma_i^1 \\ \gamma_i^2 \end{bmatrix} and\, W_i = \begin{bmatrix} W_i^1 \\ W_i^2 \end{bmatrix}$ Is $2n - vector$ of realization of a bivariate first order auto-regressive process $w_i(t) = \begin{bmatrix} w_i^1(t) \\ w_i^2(t) \end{bmatrix} and\, \varepsilon_i = \begin{bmatrix} \varepsilon_i^1 \\ \varepsilon_i^2 \end{bmatrix}$ represents independent measurement errors. The covariance matrix of measurement errors is defined by $\Sigma_i = \Sigma \otimes I_n$ and $\Sigma = \begin{bmatrix} \sigma^2_{\varepsilon^1} & 0 \\ 0 & \sigma^2_{\varepsilon^2} \end{bmatrix}$ (the symbol $\otimes$ represents the Kronecker product) . The covariance function of the bivariate auto-regressive process

$w_i(t) = \begin{bmatrix} w_i^1(t) \\ w_i^2(t) \end{bmatrix}$ is given by $R_i(s,t) = C \times e^{B|t-s|}$ with

$C = \begin{bmatrix} \sigma^2_{w^1} & \sigma_{w^1w^2} \\ \sigma_{w^1w^2} & \sigma^2_{w^2} \end{bmatrix}$ is the process covariance matrix at $t = s$ and B is a 2×2 matrix such that (i) the eigen values of B have negative parts and (ii) C and $D = -(CB + BC)$ are positive definite symmetric Sy et al(1997). The covariance matrix of random effects is the matix $G = \begin{bmatrix} G^1 & G^{12} \\ G^{12} & G^1 \end{bmatrix}$. With the assumption that $\gamma_i, W_i \text{ and } \varepsilon_i$ are mutually independent, it is obvious that

$$\text{var}(Y_i) = V_i = Z_i G_i Z_i^T + R_i + \Sigma_i \quad (4)$$

Models Using Proc MIXED SAS

Random Effects

Multivariate random effects models can be fitted using the statement random and an inductor variable for each marker to define $Y_i^k, X_i^k \text{ and } Z_i^k$. To add an independent error for each reponse variable in a multivariate random effect model, one must use the repeated statement with the option $GROUP(VAR)$ where VAR is a binary variable indicating the the response variable concerned $(VAR = 0)\ for\ Y^1$ and $VAR = 1\, Y^2$). This option allows estimation of heterogeneous covariance structure, i.e. the variances of the measurement errors are different for each response variable.

First Order Auto-regressive Process

In the repeated statement SAS provides the possibility to fit bivariate models using a Kronecker product notation Galeckit(1994) . For instance, in the bivariate case with 3 repeated measures, the option *type=UN@AR(1)* in the statement *repeated* assumes that the covariance matrix has the following structure :

$$\begin{bmatrix} \sigma^2_{w^1} & \sigma_{w^1w^2} \\ \sigma_{w^1w^2} & \sigma^2_{w^2} \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}. \quad (5)$$

Compared with the general bivariate auto-regressive process defined in the previous section, this structure has two important limitations. First, the covariance structure is a first order auto-regressive process for discrete data and assumes the measures are equally spaced for all subjects and for the two markers. In the univariate case, a continuous time AR(1) model, which allows non equally spaced measures, may be fitted using the structure *SP(POW)* but this structure is not available for multivariate models. The second limitation is that SAS program allows to estimate only one correlation parameter (ρ) for the "bivariate process" rather than a matrix B. Thus, using this formulation, one assumes that the intra-marker correlation is the same for the two markers, i.e.

$$Corr\left(w_i^1(s), w_i^1(t)\right) = Corr\left(w_i^2(s), w_i^2(t)\right) = \rho^{|t-s|} \quad (6)$$

Moreover, one assumes that inter-markercorrelation is proportional to the inter-marker correlation is proportional to the inter-marker correlation, i.e.

$Corr\left(w_i^1(s), w_i^1(t)\right) = \frac{\sigma_{w^1 w^2}}{\sigma_{w^1}\sigma_{w^2}} \rho^{|t-s|}$. Both markers are independent if the covariance matrix has the form

$$\begin{bmatrix} \sigma_{w^1}^2 & 0 \\ 0 & \sigma_{w^2}^2 \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}. \quad (7)$$

To add sn independent measurement error for both markers, one must use the option *LOAL(EXP<effect>)* which produces exponential local effects, *<effects>=VAR* being still the indicator variable of response variable. These local effects have the form

$\sigma_\varepsilon^2 diag\left[\exp(U\delta)\right]$ where U is a full–rank design matrix PROC MIXED constructs U in terms of 1s and -1s for a classification effect and estimates δ

APPLICATION

A total of 1,281 HIV-1 infected patients were enrolled from May 1997 to June 1999 at the initiation of their first highly active antiretroviral therapy containing a protease inhibitor. Standardized clinical and biological data including CD4+ cell counts measurements and plasma HIV RNA quantification were collected at baseline (M_0), one month later (M_1) and every 4 months $(M_4 - M_{24})$ thereafter. In order to ensure sufficient available information, only a sub-sample of patients having both plasma HIV RNAM and CD4+ cell counts measurements at M_0 and at least two measurements thereafter were included in the analyses. The first measurement after baseline (at one month) was deleted to provide a data with equally spaced measures. Follow-up data were included until the 24th month, thus patients had a maximum of 7 measures. Available information at each study time and description of the evolution of both markers are presented in Table 1 and Fig.1.

Modeling

To ensure normality and homoskedasticity of residuals distribution, variable response was the change in value of marker at time t since the initial visit, i.e.

$Y_i^1(t) = \log_{10} HIVRNA(t) - \log_{10} HIVRNA(0)$ and

$Y_i^2(t) = CD_4(t) - CD_4(0)$. (8)

Fixed effects included a change of slope intensity at time 4 months as suggested in fig.1. Note that we did not include intercept because

$Y_i^1(0) = Y_i^2(0) = 0 \ \forall i$. (9)

We compare 4 models providing two forms of covariance structure (random effects or auto-regressive process) in two formulations (univariate or bivariate). Univariate and bivariate random effect models were compared using likelihood ratiotest as both models were nested . The bivariate model had only four covariance parameters in addition. Comparison of random effects versus auto-regressive process were performed using AIC critrria Akaike(1974). A general model including random slopes and a bivariate first order autoregressive process did not converge as reported in univariatew cases by others (see Lesaffre et al (1999) for example). The model including two random slopes and a measurement error for each marker was

$$\begin{aligned} Y_i^1 &= \beta_i^1(t_i \wedge \tau) + \beta_2^1(t_i \wedge \tau) I_{t\geq\tau} + \gamma_{1i}^1(t_i \wedge \tau) \\ &\quad + \gamma_{2i}^1(t_i - \tau) I_{t\geq\tau} + \varepsilon_i^1 \\ Y_i^2 &= \beta_i^2(t_i \wedge \tau) + \beta_2^2(t_i \wedge \tau) I_{t\geq\tau} + \gamma_{1i}^2(t_i \wedge \tau) \\ &\quad + \gamma_{2i}^2(t_i - \tau) I_{t\geq\tau} + \varepsilon_i^2 \end{aligned} \quad (10)$$

where β_1^k is the first slope before the time $\tau = 4$ months, β_2^k is the second slope after the time τ *and* $t \wedge \tau$ represents the minimum between t_i *and* τ. Moreover,

$$\begin{pmatrix} \gamma_{1i}^1 \\ \gamma_{2i}^1 \\ \gamma_{1i}^2 \\ \gamma_{2i}^2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\gamma_{11}^1} & \sigma_{\gamma_{11}^1\gamma_{21}^1} & \sigma_{\gamma_{11}^1\gamma_{11}^2} & \sigma_{\gamma_{11}^1\gamma_{21}^2} \\ \sigma_{\gamma_{11}^1\gamma_{21}^1} & \sigma^2_{\gamma_{21}^2} & \sigma_{\gamma_{21}^1\gamma_{11}^2} & \sigma_{\gamma_{21}^1\gamma_{21}^2} \\ \sigma_{\gamma_{11}^1\gamma_{11}^2} & \sigma_{\gamma_{21}^1\gamma_{11}^1} & \sigma^2_{\gamma_{11}^2} & \sigma_{\gamma_{11}^2\gamma_{21}^2} \\ \sigma_{\gamma_{11}^1\gamma_{21}^2} & \sigma_{\gamma_{21}^1\gamma_{21}^2} & \sigma_{\gamma_{11}^2\gamma_{21}^2} & \sigma^2_{\gamma_{21}^2} \end{pmatrix} \right) \quad (11)$$

The model including an auto-regressive process and a measurement error was

$$\begin{cases} Y_i^1 = \beta_1^1 t_i \wedge \tau + \beta_2^1(t_i - \tau) I_{t\geq\tau} + W_i^1 + \varepsilon_i^1 \\ Y_i^2 = \beta_1^2 t_i \wedge \tau + \beta_2^2(t_i - \tau) I_{t\geq\tau} + W_i^2 + \varepsilon_i^2 \end{cases} \quad (12)$$

where $\begin{bmatrix} W_i^1 \\ W_i^2 \end{bmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, R_i \right)$ (13)

Data modeling using bivariate linear

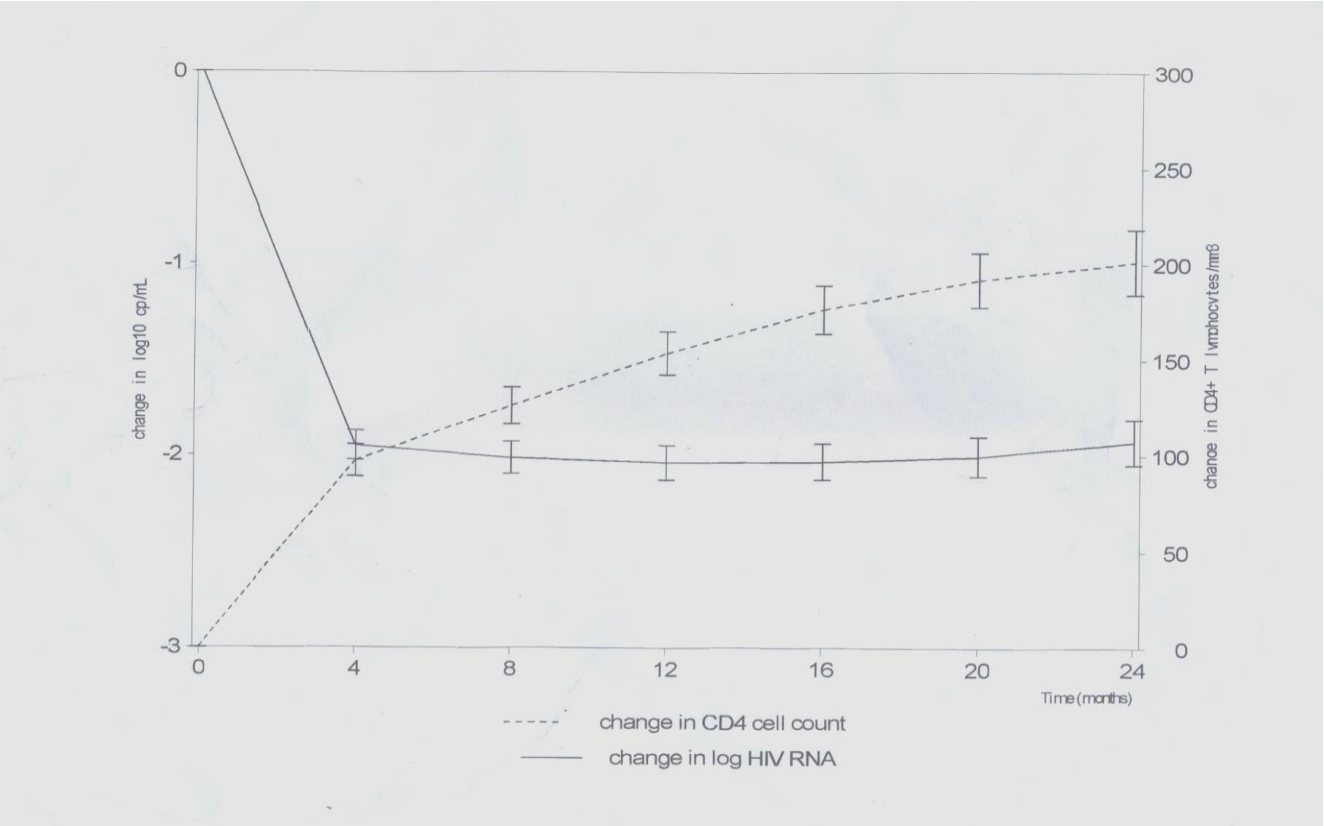

Fig.1. Mean Change in Observed HIV RNA and CD4+ Cell Count (95% Confidence Interval) after Initiation of an Antiretroviral Treatment Containing a Protease Inhibitor

Table 1. CD4 cell count and HIV RNA during follow-up

Change in CD4 cell count/ mm^3				**Change in $\log_{10}$ copies/ml HIV RNA**		
From Baseline				**From Baseline**		
	N	mean	SD	N	mean	SD
4	988	97	130	988	-1.95	1.20
8	935	126	147	919	-2.01	1.27
12	901	153	169	894	-2.04	1.34
16	823	176	180	813	-2.03	1.35
20	708	192	190	703	-2.00	1.37
24	534	201	196	530	-1.93	1.37

and

$$R_i = \begin{bmatrix} \sigma^2_{w^1} & \sigma_{w^1w^2} \\ \sigma_{w^1w^2} & \sigma^2_{w^2} \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho & \ldots & \rho^7 \\ \rho & 1 & \rho & \ldots \\ \ldots & \rho & \ldots & \rho \\ \rho^7 & \ldots & \rho & 1 \end{bmatrix} \quad (14)$$

RESULTS

The bivariate random effects model was significantly better than two separate univariate random effects models (-25194 vs -2525307, likelihood ratio = 226 with 4 degrees of freedom, $p < 10^{-4}$, (Table 2) showing a strong association between the two markers. The bivariate random effect model allows to estimate the correlation matrix between individual slopes for each marker . In this correlation matrix, every element was significantly ($p < 0.05$) different from 1 (Table 3) . Briefly, the highest correlations were between the slopes of the two markers at the same period;

$$\rho\left(\beta_1^{CD4}, \beta_1^{HIV\ RNA}\right) = -0.41\ before\ 4\ months \quad and\ \rho\left(\beta_2^{CD4}, \beta_2^{HIV\ RNA}\right) = -0.60\ after\ 4\ months \quad (15)$$

These results were expected because of biological relationship between the two markers. Moreover, the second slope of CD4 cell count was highly correlated to the first slope of the same marker $\rho\left(\beta_1^{CD4}, \beta_2^{CD4}\right) = 0.37$ The bivariate model including a bivariate auto-regressive process was better than the bivariate random effects model despite the restrictive assumption that the two intra-marker correlations are equal $(AIC\ 50386\ vs\ 50646)$. Output obtained with the model including a first order auto-regressive process provide estimations of $\sigma^2_{w^1} = 1.54, \quad \sigma^2_{w^2} = 195$ And $\sigma_{w^1w^2} = -7.00$ significantly different from 0 (wald test , $p < 10^{-4}$). This last result underlines the relationship between the two markers. The parameter $\rho = \frac{3.11}{3.42} = 0.91$ is the correlation between two consecutive measures of CD4 cell count or HIV RNA . Variances of measurement error are calculated as $\sigma^2_{\varepsilon^1} = 3{,}42e^{3.11} = 77.00$ and $\sigma^2_{\varepsilon^2} = 3.42e^{3.11} = 0.15$.

Thus, the relationship between the two markers were underlined by the correlation between the markers at each period and the improvement of likelihood of the bivariate model compared to two univariate models. Bivariate random effect model offers a direct interpretation of the relationship between the markers without assumption on the dependence of one marker in relation to the other.

CONCLUSION

Bivariate models are useful for longitudinal data in biomedical research and can be computed using standard statistical package like the SAS system. Moreover, the efficiency of the procedure MIXED, which allows quick convergence, should be underlined. However, there are some limitations inherent in the identical intra-marker correlations or the assumption of constant period between two measurement for the first order auto-regressive covariance structure implemented in the SAS system.Finally , although the number of parameters would dramatically increase, particularly in the case of multivariate random effect model, bivariate models are easily extendable to multivariate modles with more than two dependent variables.

Table 2. Likelihood of models according to the type of covariance matrix

	Log Likelihood	No. of parameters	AIC
Univariate model with two random slopes	**-25307**	**12**	**50638**
Bivariate model with two random slopes	**-25194**	**16**	**50420**
Univariate model AR(1)	**-25313**	**10**	**50646**
Bivariate model with AR(1)	**-25183**	**10**	**50386**

AIC = (-2log likelihood)+2(No. of parameter) **AR(1) : First order auto-regressive process**

Table 3. Estimated correlation matrix of the bivariate model including two random slopes

	1st Slope of HIV RNA	2nd Slope of HIV RNA	1st Slope ofCD4+	2nd Slope of CD4+
1st Slope of HIVRNA	**1**			
2nd Slope of HIV RNA	**-0.10**	**1**		
1st Slope of CD4+	**-0.41**	**0.13**	**1**	
2nd Slope of CD4+	**-0.16**	**-0.60**	**0.37**	**1**

REFERENCES

Akaike H. (1974). *A new look at the statistcal model identification.* IEEE Trans Automat Contr AC-10 :716-723

Galecki, A. T.(1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. Commun. Statist.- *Theory Meth.* 23 :3105-3119

Laird , N. M. and Ware, J. H.(1982). Random-effect models for longitudinal data. *Biometrics* 38: 963-974

Lesaffre E, Asefa M and Verbeke G (1999). Assessing the goodness-of-fit of the laird and ware model - an example : the jimma infany survival differential longitudinal study, *StatMed* 18:835-854.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithm for linearmixed- effects models for repeated-measures data, *JASA* 83: 1014-1022

Littell, R. C., Milliken, G. A. ,Stroup, W. W., and Wolfinger, R. D. (1996). *SAS System for Mixed Models.* SAS Institute, Cary, NC.

Shah, A., Laird, N. M. and Schoenfeld, D. (1997). A random-effects model for multiple characteristics with possibly missing data , *JASA* 92: 775-779

Sy, J. P., Taylor, J. M. and Cumberland, W. G. (1997). *A* stochastic model for the analysis of bivariate longitudinal AIDS data, *Biometrics* 53: 542-555

Verbeke, G. & Molenberghs, P (1997). *Linear Mixed Models in Practice.* A SAS -oriented approach. Springer, New York