Elsevier Editorial System(tm) for Computational Statistics and Data Analysis
Manuscript Draft

Manuscript Number:

Title: Conditional models with intermittent missingness: SAS code and applications

Article Type: Research Paper

Section/Category: I. Computational Statistics

Keywords: Conditional model; Likelihood; Missingness patterns; Missing data mechanism; Macro; Clinical trial

Corresponding Author: Mr. Rolando Uranga, MSc

Corresponding Author's Institution: CENCEC

First Author: Rolando Uranga, MSc

Order of Authors: Rolando Uranga, MSc; Geert Molenberghs, BS PhD

Cover Letter

RE: Conditional models with intermittent missingness: SAS code and applications, by Rolando Uranga and Geert Molenberghs.

Word Count: 7300

Dear S.P. Azen:

On behalf of my co-author, I am submitting the enclosed material for possible publication in Computational Statistics and Data Analysis (CSDA). It has not been submitted for publication nor has it been published in whole or in part elsewhere. I attest to the fact that both authors listed on the title page have read the manuscript, attest to the validity and legitimacy of the data and its interpretation, and agree to its submission to CSDA.

MCs. Rolando Uranga Piña
Responsable de Análisis Estadístico
Departamento de Diseño, Análisis y Procesamiento de Datos
Centro Nacional Coordinador de Ensayos Clínicos (CENCEC)
Telf. (537) 271 7345 Ext.119
Calle 200 Esq. 21. Atabey, Playa.
Ciudad de la Habana. Cuba.
http://www.cencec.sld.cu
rolando@cencec.sld.cu

Friday, November 26, 2010

Conditional models with intermittent missingness: SAS code and applications

R. Uranga[a,1,*], G. Molenberghs[b]

[a] *Department of Design and Data Analysis, National Center of Clinical Trials, 23 and 200 street, Atabey, Playa, Ciudad Habana, Cuba.*
[b] *Center for Statistics, Hasselt University, Agoralaan 1, B-3590 Diepenbeek, Belgium.*

Abstract

The present work provides a set of macros performed over SAS (Statistical Analysis System) for Windows, capable to fit conditional models under the problematic scenario of intermittent missingness in longitudinal data. Model fitting is based on the Missing Completely At Random (MCAR) or Missing At Random (MAR) assumptions, and the separability condition. The problem translates to maximization of the marginal observed data density only, which for Gaussian data is again Gaussian, meaning that likelihood can be expressed in terms of the mean and covariance matrix of the observed data vector, thus allowing implementation by means of a matrix oriented language like IML (Interactive Matrix Language) of SAS. A practical application is also given, where a convenient conditional model is fitted to the data from a clinical trial that assessed the effect of a Cuban product on a disease of the respiratory system. A parsimonious transition model of order seven with six parameters is obtained. A strong dependence is detected of the actual value of the primary endpoint, oxygenation index, on previous values reached one hour, three hours and seven hours before. Time distinguishes as a significant covariate: it is possible to say that oxygenation index tends to raise its values with time. This conclusion conveys a gradual improvement of patients, at least during the three days of treatment.

Keywords: Conditional model, Likelihood, Missingness patterns, Missing data mechanism, Macro, Clinical trial

[*] Corresponding author
Email address: rolando@cencec.sld.cu (R. Uranga)
[1] Tel.: 537-2717345, ext. 119. Fax: 537-2720644

1. Introduction

In a conditional model the parameters describe a feature (expectation, probability, odds, logit . . .) of a response, given values for the other responses (Cox, 1972). In a longitudinal study, the subset of conditioned responses can be conveniently selected as the set of all measurements recorded earlier in time, or maybe a subset of the more recent measurements (Molenberghs and Verbeke, 2005). Such models are known as transition models. The order of a transition model is the number of previous observations that is considered to influence the actual one. Molenberghs and Verbeke (2005) describe in detail how to implement transition models in SAS. They explain that model fitting is easy, because subsequent measurements, given their past history, are independent of each other, and hence standard software can be used, such as the SAS procedure MIXED for Gaussian data, and GENMOD and LOGISTIC for discrete data. One only needs to ensure that the previous measurement(s) can be used as a covariate and the useful macro %dropout is presented. Nevertheless, the proposed method is only valid under monotone missingness (dropout).

It is not unusual for some measurement sequences in a longitudinal study to terminate early for reasons outside the control of the investigator. Any unit so affected is called a dropout. In addition, intermediate scheduled measurements might be missed, which are termed intermittent missing values. In his 1976 paper, Rubin provides a formal framework for the field of incomplete data by introducing the important taxonomy of missing data mechanisms, consisting of missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). An MCAR mechanism potentially depends on observed covariates, but neither on observed nor unobserved outcomes. An MAR mechanism depends on the observed outcomes and perhaps also on the covariates, but not further on unobserved measurements. Finally, when an MNAR mechanism is operating, missingness does depend on unobserved measurements, maybe in addition to dependencies on covariates and/or on observed outcomes.

Rubin (1976) contributes the concept of ignorability, stating that under precise conditions, the missing data mechanism can be ignored when interest lies in inferences about the measurement process. Combined with regularity conditions, ignorability applies to MCAR and MAR combined, when likelihood or Bayesian inference routes are chosen, but the stricter MCAR condition is required for frequentist inferences to be generally valid (Rubin, 1976; Verbeke and Molenberghs, 2000; Molenberghs and Kenward, 2007).

In this work, by means of a refined use of the capabilities of SAS, valid codes under the general scenario of intermittent missingness patterns present in data collected of a longitudinal kind are provided, which allow fitting of the conditional model for Gaussian data formalized in section 2.2. These codes generalize those described in Molenberghs and Verbeke (2005). The assumptions about the missingness mechanism now relax to Missing Completely At Random (MCAR) or Missing At Random (MAR), and the validity of the regularity conditions that warrant ignorability of the missingness process. According to Rubin (1976) and Little and Rubin (1987), the problem then translates to maximization of the marginal distribution of the observed data; for Gaussian data, this marginal distribution is again Gaussian, meaning that the objective function can be expressed in terms of the mean and covariance matrix of the observed data vector. This fact allows solving the problem via a matrix oriented programming language such as IML in SAS.

In the next section some theory is revised. Section 2.1 formalizes the assumptions supporting ignorability and section 2.2 defines a general kind of transition model, with useful results. Next, in section 3, the macros are briefly described, with all the necessary code developed in the appendix. Section 4 is devoted to a practical application, where a convenient conditional model is fitted to the data from a Cuban clinical trial. A brief discussion is presented in section 5. Finally, section 6 provides some concluding remarks.

2. Theoretical framework

2.1. Direct likelihood

The present work assumes an MCAR-MAR missingness mechanism, and that the regularity conditions that warrant ignorability of the missingness process are satisfied. Namely (see Verbeke and Molenberghs, 2000), let us decide to use likelihood based estimation. The full data likelihood contribution for subject i assumes the form

$$L_i^*(\theta, \psi | y_i, r_i) \propto f(y_i, r_i | \theta, \psi)$$

Here $Y_i = (Y_{i1}, ..., Y_{in_i})$ is the response vector, $R_i = (R_{i1}, ..., R_{in_i})$ is the missingness indicator vector ($R_{ij} = 1$ if Y_{ij} is observed, $R_{ij} = 0$ if Y_{ij} is not observed), θ and ψ are parameter vectors that describe the measurement and missingness processes, respectively, and the dependence on covariates is omitted to simplify notation. Since inference has to be based on what is observed, the full data likelihood L^* has to be replaced by the observed data likelihood L:

$$L_i(\theta,\psi|y_i^o,r_i) \propto f(y_i^o,r_i|\theta,\psi)$$

with

$$f(y_i^o,r_i|\theta,\psi) = \int f(y_i,r_i|\theta,\psi)dy_i^m = \int f(y_i^o,y_i^m|\theta)f(r_i|y_i^o,y_i^m,\psi)dy_i^m$$

and

Y_i^o = subvector of observed responses

Y_i^m= subvector of not observed responses

Under a MAR process, we obtain

$$f(y_i^o,r_i|\theta,\psi) = \int f(y_i^o,y_i^m|\theta)f(r_i|y_i^o,\psi)dy_i^m = f(y_i^o|\theta)f(r_i|y_i^o,\psi)$$

If, further, θ and ψ are disjoint in the sense that the parameter space of the full vector (θ,ψ) is the product of the parameter spaces of θ and ψ, then inference can be based on the marginal observed data density only. This technical requirement is referred to as the separability condition. When the separability condition is satisfied, and within the likelihood framework, the missingness mechanism may be ignored.

2.2. General Model

A transition model, enlarged with a random effect and a measurement error, is defined next. This model constitutes a basic material for our further developments. The model is:

$$\left\{\begin{array}{l} Y_{ij} = X_{ij}\beta + b_i + \delta_{ij} + \varepsilon_{ij} \\ \delta_{ij} = \alpha_1\delta_{i,j-1} + \alpha_2\delta_{i,j-2} + ... + \alpha_n\delta_{i,j-n} + z_{ij},\, j > n \\ b_i \sim N(0,d^2); \delta_i \sim N(0,\lambda^2 H); \varepsilon_i \sim N(0,\sigma^2 I_M) \\ z_{ij}\,\text{independent of}\,\left(\delta_{\text{i1}},...,\delta_{\text{i,j}-1}\right) \\ b_1,...,b_N,\delta_1,...,\delta_N,\varepsilon_1,...,\varepsilon_N\,\text{independent} \end{array}\right. \tag{1}$$

Y_{ij} represents the response of subject i at time j, $1 \le i \le N$, $1 \le j \le M$, N is the number of subjects, M is the number of measurements over time, X_{ij} is a $(1 \times p)$ vector of known covariates, β is a $(p \times 1)$ vector of fixed effects related to the mean structure, b_i represents a random intercept, $\delta_i = (\delta_{i1}, \delta_{i2}, ..., \delta_{iM})$ is a vector of residual terms that relate the actual measurement to the previous n measurements - serial correlation component -, n is the order of the model, $z_i = (z_{i1}, z_{i2}, ..., z_{iM})$ is a vector of independent innovations and $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i1}, ..., \varepsilon_{iM})$ is an extra component reflecting the variability introduced by the measurement process. It is assumed that measurement occasions are equally spaced over time and that the correlation matrix H is of type Toeplitz - constant correlation across each diagonal -, with generic element $h_{ij} = |i - j|$, where $k = Corr(j, j + k)$, $k \geq 0$.

The model just described implies that the variance of Y_{ij} is constant and its value is $\nu^2 = Var(Y_{ij}) = d^2 + \lambda^2 + \sigma^2$. Further, the correlation between two measurements Y_{ij} and $Y_{ij'}$ only depends on the time lag $k = |j - j'|$ and its value is $r_k = Corr(Y_{ij}, Y_{ij'}) = Cov(Y_{ij}, Y_{ij'})/\nu^2 = (d^2 + \lambda^2 \rho_k)/\nu^2$. The model is sensible when there are repeated measurements over time from a certain quantitative characteristic that does not depend on subjective factors. Let α and ρ denote the column vectors defined by $\alpha = (\alpha_1, ..., \alpha_n)$, $\rho = (\rho_1, ..., \rho_n)$. Define H_m = order m matrix with generic element $h_{ij} = \rho_{|i-j|}$, and define $\alpha_j = 0$ if $j < 1$ or $j > n$. Four propositions are presented next.

Proposition 1. *The following relation holds:* $\alpha_1\rho_1 + ... + \alpha_n\rho_n + Var(z_{ij})/\lambda^2 = 1$. *In particular, the variance of* z_{ij} *is constant and will be denoted by* τ^2.

PROOF. If $j > n$ then $Corr(\delta_{ij}, \delta_{ij}) = Cov(\delta_{ij}, \delta_{ij})/\lambda^2$

$$= Cov(\alpha_1\delta_{i,j-1} + \alpha_2\delta_{i,j-2} + ... + \alpha_n\delta_{i,j-n} + z_{ij})/\lambda^2$$
$$\Rightarrow 1 = \alpha_1 Corr(\delta_{i,j-1}, \delta_{ij}) + ...\alpha_n Corr(\delta_{i,j-n}, \delta_{ij}) + Cov(z_{ij}, \delta_{ij})/\lambda^2$$
$$\Rightarrow 1 = \alpha_1\rho_1 + ...\alpha_n\rho_n + Cov(z_{ij}, \alpha_1\delta_{i,j-1} + \alpha_2\delta_{i,j-2} + ... + \alpha_n\delta_{i,j-n} + z_{ij})/\lambda^2$$
$$\Rightarrow 1 = \alpha_1\rho_1 + ...\alpha_n\rho_n + Cov(z_{ij}, z_{ij})/\lambda^2$$
$$\Rightarrow 1 = \alpha_1\rho_1 + ...\alpha_n\rho_n + Var(z_{ij})/\lambda^2$$

Proposition 2. *The following relations hold:* $\rho = H_n \cdot \alpha$, $\rho_j = \alpha_1\rho_{j-1} + \alpha_2\rho_{j-2} + +\alpha_n\rho_{j-n}$, $j \geq n$.

PROOF. If m is some natural number greater than n and j, then

$$\delta_{im} = \alpha_1\delta_{i,m-1} + ... + \alpha_n\delta_{i,m-n} + z_{im}$$

To get the first relation it is enough to take, in turn, correlation in both sides with $\delta_{i,m-1}$, $\delta_{i,m-2}$, ..., $\delta_{i,m-n}$. To get the second relation it is enough to take correlation in both sides with $\delta_{i,m-j}$.

Proposition 3. *Let A, B be the squared matrices of order n with generic elements $a_{ij} = \alpha_{i-j}$, $b_{ij} = \alpha_{i+j}$. Then $(I_n - A - B)\rho = \alpha$.*

Proof. If C is the matrix with generic element $c_{ij} = \rho_{i-j} \cdot 1_{(i>j)}$, then $H_n = I_n + C + C^T$. Also the following relations hold: $C \cdot \alpha = A \cdot \rho$, $C^T \cdot \alpha = B \cdot \rho$. Hence, $\rho = H_n \cdot \alpha = (I_n + C + C^T) \cdot \alpha = \alpha + C \cdot \alpha + C^T \cdot \alpha = \alpha + A \cdot \rho + B \cdot \rho$, and $\alpha = \rho - (A + B)\rho = (I_n - A - B)\rho$.

Proposition 4. *If the matrix H_{n+1} is positive-definite, then it is also positive-definite every H_m with $m > n$.*

Proof. Define the column vector $x = (\rho_1, \rho_2, ..., \rho_{m-1})$. Then

$$H_m = \begin{pmatrix} 1 & x^T \\ x & H_{m-1} \end{pmatrix}$$

Define the column vector $\overset{\sim}{\alpha}= (\alpha_1, \alpha_2, ..., \alpha_{m-1}) = (\alpha_1, ..., \alpha_n, 0, ..., 0)$. If in the identity

$$\delta_{im} = \alpha_1 \delta_{i,m-1} + ... + \alpha_{m-1}\delta_{i,1} + z_{im}$$

we take, in turn, correlation in both sides with $\delta_{i,m-1}$,$\delta_{i,m-2}$,...,δ_{i1}, we get

$$x = H_{m-1} \cdot \overset{\sim}{\alpha}$$

One also may write

$$\begin{aligned} \det H_m = \det \begin{pmatrix} 1 & x^T \\ x & H_{m-1} \end{pmatrix} &= \det \begin{pmatrix} 1 & x^T \\ x & H_{m-1} \end{pmatrix} \det \begin{pmatrix} 1 & 0 \\ -H_{m-1}^{-1}x & I \end{pmatrix} \\ &= \det \begin{pmatrix} 1 - x^T H_{m-1}^{-1} x & x^T \\ 0 & H_{m-1} \end{pmatrix} = (1 - x^T H_{m-1}^{-1} x) \det H_{m-1} \end{aligned}$$

The trick of multiplying by a determinant of unit value has been used.

$\therefore x^T H_{m-1}^{-1} x = x^T H_{m-1}^{-1} H_{m-1} \overset{\sim}{\alpha}= x^T \overset{\sim}{\alpha}= \rho^T \alpha \Rightarrow \det H_m = (1 - \rho^T \alpha) \det H_{m-1}$
$\Rightarrow \det H_m = (1 - \rho^T \alpha)^{m-n} \det H_n \Rightarrow \det H_{n+1} = (1 - \rho^T \alpha) \det H_n$

$\therefore H_{n+1}$ is positive-definite $\Rightarrow \det H_n > 0, \det H_{n+1} > 0$
$\Rightarrow 1 - \rho^T \alpha > 0 \Rightarrow \det H_m > 0 \Rightarrow H_m$ is positive-definite.

The propositions presented constitute a theoretical pier that justifies and guides our work. According to proposition 2, the ρ_j do not constitute new parameters, but are quantities expressible throw the α's. That's why the general model (1) depends only on $n + p + 3$ parameters $\alpha_1, ..., \alpha_n$,$\beta_1, ..., \beta_p$,d^2,λ^2,σ^2. Proposition 4 gives a necessary and sufficient condition for the correct definition of the model: model (1) is correctly defined if and only if the matrix H_{n+1} is positive-definite. This condition may be reviewed by an iterative optimization algorithm to avoid non-feasible operations; for example, extraction of squared roots to negative numbers due to the presence of matrices that in theory should be positive-definite but are not because actual estimates lie outside the parameter space. This situation will be illustrated in section 3.

3. Description of the macros

The main product of this work is developed in the appendix. It is organized in a group of macros performed over SAS 9.1.3 for Windows, which allow the fitting of the general model (1). Matrix oriented language IML of SAS is invoked. The macros are: %start, %definition, %variance, %case_i, %optimization, %estimators, %modelfitting, %dimension. Part I of SAS code in the appendix develops the first seven ones; these do not change. Macro %dimension, on the other hand, needs to be updated when model or data change. Part II of SAS code in the appendix defines a sample data set (see also section 4.1), and part III fits the data (see also section 4.3).

Macros %start and %definition generate constants and parameters; %variance creates the variance-covariance matrix of the subvector of observed data; %case_i gathers information on each individual; %optimization contains the important call to the NLPNRR subroutine of IML, which maximizes the objective function defined in module loglik of macro %modelfitting; %estimators produces estimates of parameters in the model besides precision estimates and p-values; %modelfitting produces the vector of parameter estimates besides its variance-covariance matrix. The implementation of the objective function, that is, the marginal observed data density, uses the expressions developed in propositions 1, 2, 3 of section 2.2. The feasibility of the estimates is tested via module nlc of macro %modelfitting, using the findings of proposition 4. To see results, just submit the whole SAS code (Parts I, II and III) as shown in the appendix.

4. Illustration

4.1. Cuban clinical trial

In 2004 a clinical trial on a new Cuban product intended to treat a disease of the respiratory system started. The primary endpoint was a quantitative variate known as oxygenation index, a numeric indicator of the performance of the respiratory system, with high values indicating good performance and low values, poor performance. 27 measurements of this variate were taken from each subject over time. Patients were allocated in two groups: 21 received standard therapy plus new product and 18 received standard therapy only. It was an opened, controlled, randomized phase II clinical trial. Interest focused on detection of a possible favourable effect of the new product as measured by oxygenation index, assessment of effects of relevant control variables on the response, and drawing conclusions about the evolution of the primary endpoint over time.

The product under study was administered every 8 hours during three days. Measurements of oxygenation index were collected an hour, 4 hours and 8 hours after each administration for patients who received the product, and in the scheduled times for patients in the control group. As a result, measurement occasions were (in hours after start of treatment):

$$1, 4, 8, 9, 12, 16, 17, 20, 24, 24, 28, 32, 33, 36, 40, 41, 44, 48, 49, 52, 56, 57, 60, 64,$$

$$65, 68, 72$$

Control and secondary variables were also collected. This work will focus on the control variables age, weight and height, and the important covariate group.

In the appendix (part II of SAS code), the vertically organized dataset cubanct is introduced, starting form the horizontally organized dataset cuban ct_h. Each row in cubanct_h represents a subject, and each row in cubanct represents an observation. Because there were 27 measurements of the primary endpoint per subject over time, and there are 39 subjects enrolled in the trial, cubanct has a total of 27×39 = 1053 rows. Table 1 shows a portion of the vertically organized dataset.

There are eleven variables in cubanct: *id*, *idcod*, *group*, *age*, *weight*, *height*, *day*, *evaluation*, *hour*, *time*, and *oxindex*. *id* is a numerical identifier; *idcod* is also an identifying code, with initials from the assistance institution followed by inclusion number; *group* is 1 for experimental arm and

id	idcod	group	age	weight	height	day	evaluation	hour	time	oxindex
1	AM-1	2	39	70	1.65	1	1	1	1	188
1	AM-1	2	39	70	1.65	1	1	4	4	211
1	AM-1	2	39	70	1.65	1	1	8	8	195
1	AM-1	2	39	70	1.65	1	2	1	9	153
1	AM-1	2	39	70	1.65	1	2	4	12	233
1	AM-1	2	39	70	1.65	1	2	8	16	·
	. . .		. . .				. . .			. . .
1	AM-1	2	39	70	1.65	3	3	1	65	·
1	AM-1	2	39	70	1.65	3	3	4	68	·
1	AM-1	2	39	70	1.65	3	3	8	72	·
	. . .		. . .				. . .			. . .
39	VIL-2	2	38	65	1.68	1	1	1	1	132
39	VIL-2	2	38	65	1.68	1	1	4	4	·
39	VIL-2	2	38	65	1.68	1	1	8	8	143
39	VIL-2	2	38	65	1.68	1	2	1	9	·
39	VIL-2	2	38	65	1.68	1	2	4	12	·
39	VIL-2	2	38	65	1.68	1	2	8	16	·
	. . .		. . .				. . .			. . .
39	VIL-2	2	38	65	1.68	3	3	8	72	·

Table 1: Portion of cubanct dataset.

2 for control arm; *age* (years), *weight* (kg) and *height* (m) describe general characteristics; *day* (1, 2, 3), *evaluation* (1, 2, 3), *hour* (1, 4, 8), and *time* (from 1 to 72 hours) describe each measurement occasion (for example: *day* 2, *evaluation* 3, *hour* 4 will always imply $time = 24 + 16 + 4 = 44$ hours, wich means that a measurement of the primary endpoint was collected at hour 4 of evaluation 3 of day 2, that is, 44 hours after start of treatment); finally, *oxindex* records values collected from oxygenation index.

Next, the enlarged dataset cubanctiml is defined. For each subject, there are 72 rows which represent values of corresponding variables, hour after hour, from hour 1 to hour 72: *oxindex* is filled with missing data - empty cells - in new occasions, because its value is unknown; *time* takes on all values from 1 to 72; *id*, *idcod*, *age*, *weight* and *height* remain constant per subject. New variables are added: *logoxindex* (natural logarithm of oxygenation index) and $group1$ (indicator variable defined as 1 if $group = 1$, and 0 if $group = 2$).

4.2. Exploratory data analysis

Figure 1 shows logarithmic oxygenation index mean observed profiles per treatment group over time. Initially both profiles overlap, but towards the end a favorable trend for the experimental group is perceived, with systematic higher values of the primary endpoint. The fitting of a convenient model will confirm or disclaim this assertion.

Figure 2 shows squared ordinary least squares (OLS) residuals from a model with a saturated mean structure (all available covariates plus the group-by-time interaction), thus ignoring that not all measurements are independent, and a smoothed average trend. This graph, together with plots of OLS residual profiles (not shown), allows assuming a constant variance over time.

Sampling correlations of standardized residuals, for pairs of time points 1 hour apart, may be calculated. Their value is around 0.74. Similar calculations may be performed for correlations of residuals 3 hours apart, 4 hours apart... Preliminary results are compatible with the assumption that correlations of logarithmic oxygenation index depend on the time lag only.

Moreover, following Verbeke and Molenberghs (2000), one can summarize each individual response vector by a linear combination of its components, standardize these linear combinations, and then apply a usual normality test to these summaries, calculated by replacing all parameters by their maximum likelihood estimates. As an example, the Shapiro-Wilk test does not reject the null hypothesis of normality, when applied to the standardized sums of

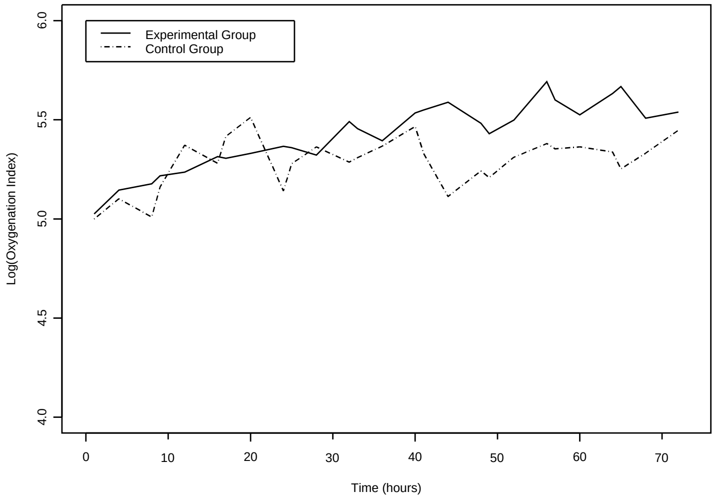

Figure 1: Mean profiles of transformed primary endpoint per group.

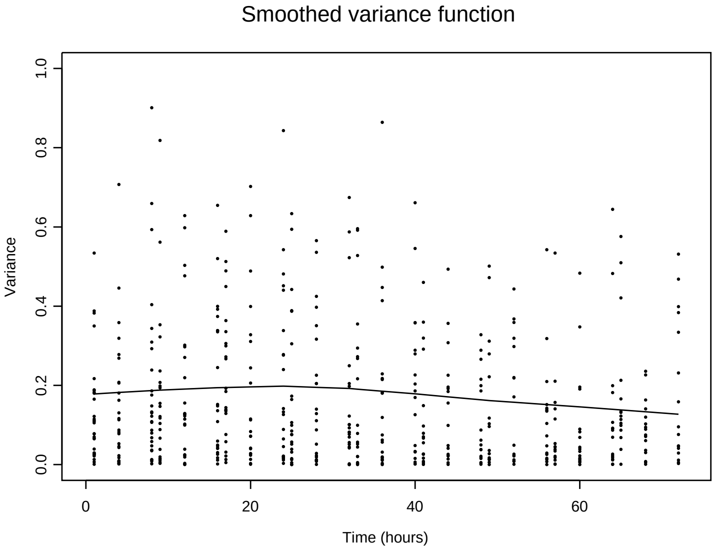

Figure 2: Smoothed average trend of squared OLS residuals. Squared residuals larger than 1 are not shown.

the components of the individual responses (logarithmic oxygenation index) for the Cuban data, giving a p-value of 0.49. Histograms and normality plots confirm this behaviour.

4.3. Model fitting

The results of the preceding section allow conceiving a model of the general kind (1) as a reasonable candidate for fitting the Cuban data. Tree facts support this decision: the approximate constancy of the variance function, the probable dependency of correlations on the time lag only, and the consistency of the normality assumption with the behaviour of the transformed primary endpoint.

Hence, as a first candidate for fitting the Cuban data, an order-eighth transition model is considered. This initial model includes all available covariates age, weight, height, group, time, and also the group-by-time interaction. The dependent variable is logarithmic oxygenation index. A heuristic justification for the choice of the eighth order is that, because the new product was administered each 8 hours, it is reasonable to hope it be unnecessary to rely on previous measurements more than 8 hours apart, to explain the actual measurement of the primary endpoint. Thus our initial, saturated model is model (1) with $X_{ij} = (1, group_i, age_i, weight_i, height_i, time_{ij}, group_i{\cdot}time_{ij})$, $n = 8, p = 7, N = 39, M = 72$, and with $group_i$ taking on the value of the indicator variable *group*1 in the i-th patient - see section 4.1 -. When a backward selection method is applied, the following reduced model is obtained, with *beta*-parameters renowned:

$$\begin{cases} Y_{ij} = \beta_1 + \beta_2 time_{ij} + \delta_{ij} \\ \delta_{ij} = \alpha_1\delta_{i,j-1} + \alpha_3\delta_{i,j-3} + \alpha_7\delta_{i,j-7} + z_{ij},\ j > 7 \\ \delta_i \sim N(0, \lambda^2 H);\, z_{ij}\text{ independent of }(\delta_{\mathrm{i1}}, ..., \delta_{\mathrm{i,j-1}}) \\ \delta_1, ..., \delta_N, \varepsilon_1, ..., \varepsilon_N\text{ independent} \end{cases} \tag{2}$$

This is model (1) with six parameters given by β_1, β_2, α_1, α_3, α_7, λ^2, and assuming $X_{ij} = (1, time_{ij})$, $p = 2$, $n = 7$, $N = 39$, $M = 72$, $\alpha_2 = \alpha_4 = \alpha_5 = \alpha_6 = d^2 = \sigma^2 = 0$. It is a classical transition model of order seven.

Part III of SAS code in the appendix performs model fitting. Stacked design matrix x, stacked response vector y, preliminary parameter vector *initial*, and constants *nsub* (number of subjects) and *ntime* (number of measurements per subject), are first created via PROC IML. Next, in macro %dimension, values are assigned to the following constants: *nbeta* = number of

Effect	Parameter	Estimate (s.e.)
Mean structure:		
Intercept	β_1	5.096 (0.065)
Time	β_2	0.007 (0.001)
Recurrent parameters:		
One hour before	α_1	0.404 (0.050)
Three hours before	α_3	0.274 (0.051)
Seven hours before	α_7	0.247 (0.040)
Variance:		
var(δ_{ij})	λ^2	0.192 (0.023)

Table 2: Parameter estimates and standard errors obtained from fitting the reduced model (2).

parameters β (p in model (1)); $nalpha$ = order of the model (n); $palpha0$ = components of α to be estimated (the remaining components are set equal to 0); $nalfa0$ = number of components to be estimated from α; nd = indicator for d^2 ($nd = 0$ if d^2 is set equal to 0; $nd = 1$ otherwise); $nlambda$ = indicator for λ^2 ($nlambda = 0$ if λ^2 is set equal to 0; $nlambda = 1$ otherwise); $nsigma$ = indicator for σ^2 ($nsigma = 0$ if σ^2 is set equal to 0; $nsigma = 1$ otherwise); $npar$ = number of parameters. Finally, macros %modelfitting and %estimators are invoked.

Table 2 presents results from fitting the reduced, final model (2). Only time remains as a significant covariate and a strong dependence is detected of the actual value of oxygenation index on previous values reached one hour, three hours and seven hours before. According to these results, no evidence is detected of an obvious effect of the new product under study on the behaviour of the primary endpoint. Neither an influence is detected of covariates age, weight or height, but we can do claim that oxygenation index tends to raise its values with the course of time, at least during the 72 hours of treatment; that is, the condition of the patient improves. Namely, after each hour, the primary endpoint oxygenation index raises its values by an amount of 0.7%.

5. Discussion

Figure 3 shows observed and fitted individual profiles of logarithmic oxygenation index according to model (2), under a conditional interpretation.

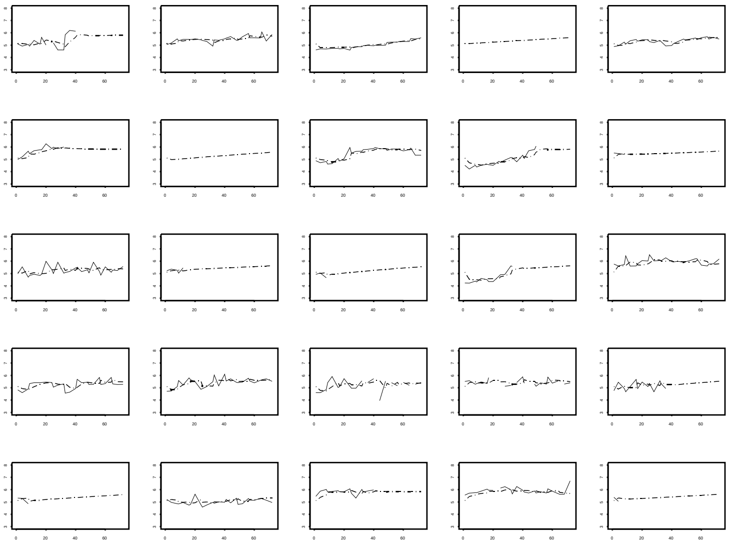

Figure 3: Conditional fitting (dashed lines): Irregularity is recognized.

That is, each point in the profile is defined as the expected actual response, given the past observed responses. This interpretation is an advantage of the conditional model formulation. It should be emphasized that, under general model (1), the response vector $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{iM})$ is multivariate normal with mean $X_i\beta$, where X_i is the matrix with i-th row equal to X_{ij}, and variance-covariance matrix $V = d^2 J_M + \lambda^2 H + \sigma^2 I_M$, where J_M is an $M \times M$ matrix of ones. Hence, under a marginal interpretation, the fitted profiles implied by model (2) are merely straight lines. The conditional interpretation allows keeping a record of the past history, thereby recognizing and absorbing irregularities in the profiles. The conditional formulation provides a parsimonious and elegant description of the driving mechanism behind such irregularities. Further, the parameters in model (1) admit a marginal interpretation, allowing the study of average characteristics in populations, for example group effects in controlled clinical trials.

It could be confusing the claim that no obvious evidence is detected of an effect of the Cuban product on one side, and the claim that oxygenation index tends to raise its values with the course of time on the other side. The first fact is a between-groups effect, and the second a within-group effect. Thus, it should be no contradiction with the absence of one, and the presence of the other. An explanation of the absence of a group effect and, nevertheless, the presence of a time effect in the Cuban data, is a probable gradual response of subjects to standard therapy.

This work is based on the MCAR-MAR assumption. As stated by Molenberghs and Verbeke (2005), it is difficult to exclude the option of a more general missingness mechanism. One solution is to fit an MNAR model. However, as pointed out by several authors, one has to be extremely careful with interpreting evidence for or against MNAR using only the data under study. A sensible compromise between blindly shifting to MNAR models or ignoring them altogether is to make them a component of a sensitivity analysis.

In sensitivity analysis, several statistical models are considered simultaneously and/or a statistical model is further scrutinized using specialized tools, such as diagnostic measures. One such promising tool, proposed by Verbeke et al. (2001), and employed by Thijs et al. (2000) and Molenberghs et al. (2001), is based on local influence (Cook, 1986). Another option is to consider pattern-mixture models as a complement to selection models (Thijs et al., 2002; Michiels et al., 2002). Important concepts to consider are: imprecision, resulting from the stochastic component of the model and for finite sampling, and uncertainty, arising from incompleteness in the data. Both can be combined into uncertainty (Kenward et al., 2001).

It has been considered in this work the general model (1), where the value of the actual error term is expressed in terms of previous error terms. Alternatively, the value of the actual outcome may be expressed in terms of values of previous outcomes. Reasoning and code are easily translated to the second situation, but as Diggle et al. (1994) point out, coefficients then loose their marginal interpretation. An advantage of the second situation is the simplicity of codes when there is not random effect or measurement error (usual transition model). In such a case, if missingness patterns are only monotone, a standard code as described by Molenberghs and Verbeke (2005) may be used, resting on PROC MIXED. In the presence of intermittent missingness, it is enough to rest on Multiple Imputation to monotonize patterns, and then to apply PROC MIXED.

We have provided codes for direct implementation of the likelihood via PROC IML of SAS. An alternative approach is to rest on the Multiple Imputation method. Thus, intermittent missingness patterns are previously transformed to monotone and the problem of intermittence disappears. Hence, in the absence of a random effect or a measurement error, standard code may be used. However, in the presence of either one of these two stochastic components, it is necessary to rely on IML macros or, as an alternative, on PROC NLMIXED.

The question may arise about the correct definition of general model (1), which includes a measurement error ε_{ij} when there are already independent innovations z_{ij}. The answer is that the variance parameter τ^2 of the z_{ij}, is related to λ^2, that is, it is not a free parameter. The inclusion of an error term pursues to add the free parameter σ^2, thus giving greater flexibility to the model.

The main products of this work are the IML macros. There were difficulties in the way that were necessary to overcome. For example, in a beginning the blc argument of the NLPNRR optimization subroutine was invoked to avoid interruption of the iterative process by error messages (squared roots of negative numbers, non-feasible solutions...). The alternative subroutine NLPQN was then invoked, which admits the "nlc" argument (non-linear constrains), with equally adverse results because the implemented algorithm is not a "feasible point algorithm" (see SAS Institute Inc., 2004). Finally, NLPNRR was again invoked and the problem was solved though the resource of assigning a missing data to the likelihood value, in case of violation of the feasibility conditions; this trick forces the subroutine to ignore the actual guess and to test a different one.

With respect to departures from normality, a mixture of normals could be assumed rather than a normal distribution. But in that case, we suggest relying on the EM (Expectation-Maximization) algorithm, described in Dempster et al. (1977), for likelihood implementation. The adaptation of the SAS/IML macros is a matter of future developments.

Morariu and Buimaga-Iarinca (2009) report an application of autoregressive modelling on coding sequence lengths in bacterial genome. Ding et al. (2007) apply autoregressive modelling to electrocardiogram data. Other fields of application of autoregressive modelling are: agriculture (annual crop yield of sugar-beets and their price per ton for example), business (daily stock prices, weekly interest rates, monthly rates of unemployment and annual turnovers), meteorology (hourly wind speeds, daily maximum and minimum temperatures and annual rainfall), geophysics (continuously observation of the shaking or trembling of the earth in order to predict possibly impending earthquakes), medicine (an electroencephalogram traces brain waves made by an electroencephalograph in order to detect a cerebral disease, an electrocardiogram traces heart waves), social sciences (survey of annual death and birth rates, number of accidents in the home and various forms of criminal activities), and quality assurance (parameters in a manufacturing process are permanently monitored in order to carry out on-line inspections).

In general, for data recorded longitudinally in time, it is always recommended to consider the option of a conditional modelling and interpretation to gain a better understanding of the data generating mechanism or to predict future values of the response. The assumption of linear dependence serves to simplify, or make possible, a theoretical analysis.

6. Concluding remarks

We may conclude that it is possible to fit transition models in the presence of intermittent missingness. This work provides a set of SAS/IML macros to this respect. A transition model of order 7 with few parameters was fitted to the Cuban clinical trial data. No evidence of an effect of the new product was detected, nor an effect of the control variables age, weight or height. A time effect was nevertheless detected. We can state that the status of the patient tends to improve with the course of time, probable due to a gradual response to standard therapy. It is recommended to apply the results obtained in this work to real situations, with the objective of accumulating practical experience and to evaluate the behaviour of the presented tools. Also, not to give inappropriate weight to the conclusions obtained about the Cuban clinical trial data. Only when other techniques be applied that confirm or disclaim our results, conclusions about, say, the effect of the Cuban product under study should be drawn.

Acknowledgements

The authors would like to thank the National Center of Clinical Trials, Cuba, and the Center for Statistics (CenStat), Universiteit Hasselt, Belgium, for support of this work.

References

R.D. Cook, Assessment of local influence, Journal of the Royal Statistical Society, Series B, **48**, pp. 133-169.

D.R. Cox, The analysis of multivariate binary data, Applied Statistics **21** (1972), pp. 113-120.

A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), Journal of the Royal Statistical Society, Series B, **39** (1977), pp. 1-38.

P.J. Diggle, K.-Y.Liang and S.L. Zeger, Analysis of Longitudinal Data, Oxford Science Publications, Oxford: Clarendon Press (1994).

G.E. Ding-Fei, H. Bei-Ping, X. Xin-Jian, Study of Feature Extraction Based on Autoregressive Modeling in ECG Automatic Diagnosis, Acta Automatica Sinica, Vol. 33, No. 5 (2007).

M.G. Kenward, E.J.T. Goetghebeur and G. Molenberghs, Sensitivity analysis of incomplete categorical data, Statistical Modelling, **1** (2001), pp. 31-48.

R.J.A. Little and D.B. Rubin, Statistical Analysis with Missing Data, New York: John Wiley & Sons, (1987).

B. Michiels, G. Molenberghs, L. Bijnens and T. Vangeneugden, Selection models and pattern-mixture models to analyze longitudinal quality of life data subject to dropout, Statistics in Medicine, **21** (2002), pp. 1023-1041.

G. Molenberghs and M.G. Kenward, Missing Data in Clinical Studies, John Wiley & Sons Ltd, Chichester, UK (2007).

G. Molenberghs and G. Verbeke, Models for discrete longitudinal data, New York: Springer - Verlag (2005).

G. Molenberghs, G. Verbeke, H. Thijs, E. Lesaffre and M.G. Kenward, Mastitis in dairy cattle: influence analysis to assess sensitivity of the dropout process, Computational Statistics and Data Analysis, **37** (2001), pp. 93-113.

V.V. Morariu, L. Buimaga-Iarinca, Autoregresive modeling of coding sequence lengths in bacterial genome, arXiv Preprint archive [on line], http://arxiv.org/ftp/arxiv/papers/0907/0907.1159.pdf (2009).

D.B. Rubin, Inference and missing data, Biometrika, **63** (1976), pp. 581-592.

SAS Institute Inc., SAS OnlineDoc® 9.1.3, Cary, NC: SAS Institute Inc. (2004).

H. Thijs, G. Molenberghs, B. Michiels, G. Verbeke and D. Curran, Strategies to fit pattern-mixture models, Biostatistics, **3** (2002), pp. 245-265.

H. Thijs, G. Molenberghs and G. Verbeke (2000), The milk protein trial: influence analysis of the dropout process, Biometrical Journal, **42** (2000), 617-646.

G. Verbeke and G. Molenberghs, Linear Mixed Models for Longitudinal Data, New York: Springer - Verlag (2000).

G. Verbeke, G. Molenberghs, H. Thijs, E. Lesaffre and M.G. Kenward, Sensitivity analysis for non-random dropout: a local influence approach, Biometrics, **57** (2001), 7-14.

Appendix

```
SAS Code

/*Part I: SAS macros*/

%macro start;use x;read all into x;use y;read all into y;
use nsub;read all into nsub;use ntime;read all into ntime;
use initial;read all into initial;%mend;
%macro definition;beta=parameters[1:nbeta];
alpha0=parameters[nbeta+1:nbeta+nalpha0];
alpha=j(nalpha,1,0);alpha[palpha0]=alpha0;d2=0;lambda2=0;
sigma2=0;if nd then d2=parameters[nbeta+nalpha0+nd];
if nlambda then lambda2=parameters[nbeta+nalpha0+nd+nlambda];
if nsigma then sigma2=parameters[npar];alphamat=i(nalpha);
do i=1 to nalpha;do j=1 to nalpha;
if i+j<=nalpha then alphamat[i,j]=alphamat[i,j]-alpha[i+j];
if i>j then alphamat[i,j]=alphamat[i,j]-alpha[i-j];end;end;
rho=inv(alphamat)*alpha;nrho=nalpha;nrhho=max(nrho,nrhho);
rhho=j(nrhho,1,0);rhho[1:nrho]=rho;do j=nrho+1 to nrhho;
```

```
rhho[j]=t(alpha)*rhho[j-1:j-nrho];end;%mend;
%macro variance;
vcciobs=lambda2*toeplitz(1//rhho)[indexobs,indexobs];
viobs=d2+vcciobs+sigma2*i(dimyiobs);%mend;
%macro case_i;xi=x[(ind-1)*ntime+1:ind*ntime,];
yi=y[(ind-1)*ntime+1:ind*ntime];mui=xi*beta;
pi=constant("PI");dimyiobs=ncol(loc(yi^=.));
if dimyiobs>0 then do;indexobs=t(loc(yi^=.));
xiobs=xi[indexobs,];yiobs=yi[indexobs];muiobs=mui[indexobs];
%variance;invviobs=inv(viobs);detviobs=det(viobs);end;%mend;
%macro optimization;
con=j(2,npar,.);opt=j(1,11,.);opt[1]=1;opt[2]=5;
call nlpnrr(rc,parest,"loglik",initial,opt,con);%mend;
%macro estimators;proc iml;%start;%dimension;use parest;
read all into parest;use covar;read all into covar;
var=vecdiag(covar);stde=sqrt(var);start cor_1(parameters);
nrhho=1;%dimension;%definition;var_y=d2+lambda2+sigma2;
cor_1=(lambda2*rhho[1]+d2)/var_y;return(cor_1);finish cor_1;
call nlpfdd(cor_1,grad_1,hessian_1,"cor_1",parest);
var_cor_1=grad_1*covar*t(grad_1);stde_cor_1=sqrt(var_cor_1);
parest=parest//cor_1;stde=stde//stde_cor_1;zval=parest/stde;
pval=2*(1-cdf("normal",abs(zval),0,1));pval[nbeta+nalpha0+1:
npar]=1-cdf("normal",zval[nbeta+nalpha0+1:npar],0,1);
create result var {parest stde zval pval};append;
namesbeta=concat("beta",compress(char(1:nbeta)));
namesalpha=concat("alpha",compress(char(t(palpha0))));
namesvar=concat("var_",compress(char(1:nd+nlambda+nsigma)));
namescor="cor_1";
parameter=t(namesbeta||namesalpha||namesvar||namescor);
create parameters from parameter[colname="parameter"];
append from parameter;data result;merge parameters result;run;
proc print data=result;run;quit;%mend;
%macro modelfitting;
proc iml;start nlc(parameters,rho,nrho);c=j(nrho,1,0);
do i=1 to nrho;c[i]=det(toeplitz(1//rho[1:i]));end;
return(c);finish nlc;
start loglik(parameters) global(x,y,nsub,ntime,nrun);
nrhho=ntime;%dimension;%definition;
```

```
if nlc(parameters,rho,nrho)>0 & d2+lambda2+sigma2>0 then do;
ll=0;ind=1;do while (ind<=nsub);%case_i;lli=0;if dimyiobs>0
then lli=-0.5*dimyiobs*log(2*pi)-0.5*log(detviobs)
-0.5*t(yiobs-muiobs)*invviobs*(yiobs-muiobs);
ll=ll+lli;ind=ind+1;end;loglik=ll;nrun=nrun+1;file log;
put nrun;end;else loglik=.;return(loglik);finish loglik;
nrun=0;nrungrad=0;%start;%dimension;%optimization;
call nlpfdd(maxlik,grad,hessian,"loglik",parest);
inf=-hessian;covar=inv(inf);var=vecdiag(covar);stde=sqrt(var);
create parest var {parest};append;
create covar from covar;append from covar;quit;%mend;

/*Part II: Code to generate data sets*/

proc datasets kill;quit;
/*Horizontal Data Set*/
data cubanct_h;input id idcod $5. group age weight height
d1e1h1 d1e1h4 d1e1h8 d1e2h1 d1e2h4 d1e2h8 d1e3h1 d1e3h4 d1e3h8
d2e1h1 d2e1h4 d2e1h8 d2e2h1 d2e2h4 d2e2h8 d2e3h1 d2e3h4 d2e3h8
d3e1h1 d3e1h4 d3e1h8 d3e2h1 d3e2h4 d3e2h8 d3e3h1 d3e3h4 d3e3h8
@@;datalines;
1 AM-1 2 39 70 1.65 188 211 195 153 233 . . . . . . . . . . .
. . . . . . . . . . . 2 AM-10 1 47 135 1.78 148 212 160 226
202 205 313 224 200 184 205 190 190 150 204 230 233 210 . .
. . . . . . . 3 AM-11 2 27 105 1.72 104.5 96.1 86.3 104.6 103
90.1 91.1 94.6 100.1 91.1 80.6 109.8 100.6 109.3 102.8 89.5
88 55.6 54.6 104.9 126.6 222.9 221.6 83.5 91.5 115.8 153.2
4 AM-2 1 57 63 1.73 132 141 190 162 213 233 205 228.8 229.7
197.3 183.5 217.1 194.8 139 144 165 197.4 244 228.4 239 266.2
230 272.4 292.8 243 278 242 5 AM-3 1 61 70 1.6 183 144 129
127 143 115 122.5 280.1 120 97.4 116.9 144 134 150 146 182.5
138 202.5 121.8 129.2 196.5 174 169 199 193 171 142 6 AM-4 2
62 70 1.7 144.2 157.8 . . . 139.2 348 . . 239 . . . . . . . .
. . . . . . . . . 7 AM-5 2 42 80 . 163 136.4 157.6 137 217
162 280 149 . 178 100 100 345 487 456 . . . . . . . . . . . .
8 AM-6 2 60 65 1.62 146.9 181.8 286 236.8 298 320 320 522 354
395 353.2 400 . . . . . . . . . . . . . . . 9 AM-7 1 21 70
1.75 234 358 408 336 350 371 340 340 433 302 203 411 330 361
```

```
396 . . . . . . . . . . . . . 10 AM-8 1 33 95 1.7 240 297 276
254 262 353 328 302 280 296 280 . . . . . . . . . . . . . . . .
. 11 AM-9 1 54 90 1.65 91 68 91 80 91 100 96 91 127 116 140
173 171 121 208 161 300 331 433 . . . . . . . . 12 CC-1 1 58
75 1.65 99.4 170.8 . . . . . . . . . . . . . . . . . . . . . .
. . . . 13 CG-1 2 32 60 1.6 100 110 108 110 114 114 110 114
99 123 130 130 143 146 140 148 146 148 185 190 190 200 200
200 253 244 270 14 CG-2 1 38 95 1.76 203 200 129 . . . . . .
. . . . . . . . . . . . . . . . . . . 15 CJF-1 1 58 56.5 1.73
320 276 308 629.9 272 275.6 326.1 424 410 684.2 417.5 423 427
532.1 390.9 380.2 399.2 376.3 383.4 421.5 491.1 491 293 273.8
321.2 328.3 483.9 16 CJF-2 2 67 60 1.57 112 114 162 267 186
331 285 244 131 138 164 241 417 173 442 256 304 229 224 235
322 257 222 270 285 305 251 17 CJF-3 1 75 54 1.63 156 176.5
246 213 237 236 233 247 229 215.8 198.5 137.5 226 219 245 261
297 219 219 294 378.4 267.7 266.8 266.2 433 206 353 18 CJF-5
1 44 50 1.55 132 113.3 125 101 104.7 157.8 128.7 155.3 388.2
244.8 284 288 325 335 360.2 388.9 349.2 353.3 304.6 343.5 343
328 293 326.4 367.4 208.9 207.6 19 CJF-6 2 29 70 1.78 218.3
158 . . . . . . . . . . . . . . . . . . . . . . . . .
20 CJF-7 1 36 50 1.6 117.1 231.7 139.8 107.5 163.2 289.2
139.4 235.8 163.2 209.7 107.5 261.7 202.3 139.8 . . . . . . .
. . . . . . 21 CJF-8 2 72 65 1.63 100 100 128 224 370 153 162
310 167.6 143 143.4 259 . 226 311.5 . 52 252.8 . 223.6 173.9
. 229 181.1 . 196.9 225.4 22 EG-1 1 46 75 . 68.6 68.12 79.8
75 99.8 89.8 77.1 77.1 120 136 139 274 265 . . . . . . . . .
. . . . . 23 EG-4 2 18 70 . 259 303 316 316 356 418 386 378 .
457 515.5 399 284.2 516.8 383 331 306.7 363.2 306.7 341.5 307
431.5 356.5 288 276 278.3 826.6 24 EG-5 2 18 57 1.67 162.16
274.6 255.6 258.3 247.7 209.7 210.8 255.6 119.4 106.94 216.9
198.8 90 119.9 107.7 194.7 230 122.11 176.4 99.05 159.4 143.8
104.1 168.4 76.35 169 80 25 HA-1 1 52 90 1.7 170 232 254 266
280 296 130 230 144 197 254 292 280 268 240 272 276 156 200
220 182 292 334 356 380 380 306 26 HA-2 2 53 80 1.73 . . . .
. . . . . . . . . . . . . . . . . . . . . . . . 27 HA-3 1 66 80
1.7 134.4 144.2 143 122.6 142.8 167 124.4 151.6 161 244 138
226 241 181 198 203 156 159 156 159 323 323 197 174 198 230
175.8 28 HA-4 2 39 100 1.82 180 112.8 126 145.2 141.6 185.3
```

```
220 198.8 147 258 310 126 200 194 141.8 150.6 115 169.3 130.8
99.8 208 116.2 92.4 214 130 137 102.6 29 LDS-1 2 41 65 .
170.1 205.8 149.9 . . . . . . . . . . . . . . . . . . . . . .
. . 30 LDS-2 1 65 . . 121.6 143.2 283.9 414 176 136 378.2
157.3 243.2 407 345.2 342.2 396.8 418.8 395 385.75 393.4
387.5 377.7 321.5 330.8 142 257.5 375.75 371.2 304.8 345
31 LDS-3 2 20 52 . 145.1 254.3 111.3 126.6 140.1 126.6 140.1
401.8 201.1 149.7 371.7 153.8 161.4 175.5 233.4 243.3 174.1
201.9 156.4 369.1 175.6 129.5 256.8 164.8 197.6 189 262.1
32 MA-1 1 35 100 1.75 124 100 134 208 224 224 226 236 226 156
184 204 96 104 140 296 224 232 192 196 342 192 208 342 204
192 196 33 ME-1 1 38 100 . 128 244 218.6 206.8 282 363.4 256
. 191.2 183.4 404.3 472 383.6 386 401 400.1 441.6 360 . 364.2
489.2 478 405 464 446.2 415 372 34 ME-2 1 58 65 . 247 263
198.8 206.6 239.4 210.6 336 . 225 . 168 180 . 226 359 217 .
199.6 167.4 221.8 201.2 349.4 216.2 261 . 199.8 222.4 35 ME-3
2 45 200 1.65 144 233.4 100.4 153 464 431.2 465 437.8 460 458
373 329 329 318 444 420 337.6 426 496 442 446 370 462 518 496
334 470 36 ME-4 2 74 75 1.74 138 155.8 104.8 . . . . . . . .
. . . . . . . . . . . . . . . . 37 SA-1 1 40 99.45 1.65 130
124 129.2 147.4 141.6 159 286 214 268 175 216 176 174 220 246
234 270 134.2 169.8 195 212 208 191 170 195 177.8 200
38 VIL-1 1 66 65 1.6 246 238 220 . . . . . . . . . . . . . .
. . . . . . . . . . 39 VIL-2 2 38 65 1.68 132 . 143 . . . . .
. . . . . . . . . . . . . . . . . . .
;
/*Vertical Data Set*/
data cubanct;set cubanct_h;array y(27) d1e1h1--d3e3h8;
do j=1 to 27;day=int((j-1)/9+1);evaluation=mod(int((j-1)/3),3)
+1;hour=(mod(j,3)=1)+4*(mod(j,3)=2)+8*(mod(j,3)=0);time=24
*(day=2)+48*(day=3)+8*(evaluation=2)+16*(evaluation=3)+hour;
oxindex=y(j);output;end;keep id idcod group age weight height
day evaluation hour time oxindex;run;
/*IML Data Set*/
data cubanctiml;set cubanct_h;keep id idcod group age weight
height;run;data cubanctiml;set cubanctiml;do time=1 to 72;
output;end;run;data cubanctiml;merge cubanct cubanctiml;
by idcod time;logoxindex=log(oxindex);group1=group=1;run;
```

```
/*Part III: Code for model fitting*/

proc iml;use cubanctiml;read all var {time logoxindex} into
data;intercept=j(nrow(data),1,1);time=data[,1];create x var
{intercept time};append;y=data[,2];create y var {y};append;
beta=4//0;alpha0=0//0//0;lambda2=0.04;initial=beta//alpha0//
lambda2;create initial var{initial};append;nsub=39;create
nsub var {nsub};append;ntime=72;create ntime var {ntime};
append;quit;%macro dimension;nbeta=2;nalpha=7;palpha0=1//3//7;
nalpha0=nrow(palpha0);nd=0;nlambda=1;nsigma=0;npar=nbeta
+nalpha0+nd+nlambda+nsigma;%mend;%modelfitting;%estimators;
```