

# Estimación de las componentes de un modelo de coeficientes dinámicos mediante las ecuaciones de estimación generalizadas

## Time-Varying Coefficient Model Component Estimation Through Generalized Estimation Equations

JUAN CAMILO SOSA<sup>a</sup>, LUIS GUILLERMO DÍAZ<sup>b</sup>

DEPARTAMENTO DE ESTADÍSTICA, FACULTAD DE CIENCIAS, UNIVERSIDAD NACIONAL DE COLOMBIA, BOGOTÁ, COLOMBIA

---

### Resumen

Se propone una metodología para estimar las componentes de un modelo de coeficientes dinámicos mediante las ecuaciones de estimación generalizadas (Liang & Zeger 1986), con el propósito de incluir directamente en la estimación la posible correlación de las medidas repetidas de cada individuo. La expansión de los coeficientes dinámicos del modelo se lleva a cabo a través de regresión *spline* (Huang et al. 2002). También se propone utilizar el *criterio de información de Akaike en las ecuaciones de estimación generalizadas* (QIC) (Pan 2001) como selector de modelos. Mediante simulación se compara la metodología propuesta con la metodología presentada por Wu & Zhang (2006), donde se estiman las componentes del modelo mediante mínimos cuadrados ponderados y se utiliza el *criterio de información de Akaike* (AIC) como selector de modelos, obteniéndose que en los escenarios simulados la metodología propuesta presenta mejores resultados con relación al error cuadrático medio promedio. Para ilustrar la estrategia de estimación propuesta, se considera el conjunto de datos ACTG 315 (Liang et al. 2003) asociado con un estudio de sida, en el que se modela dinámicamente la relación entre la carga viral y el conteo de células CD4+.

**Palabras clave:** criterio de información de Akaike, ecuaciones de estimación generalizadas, mínimos cuadrados ponderados, modelo de coeficientes dinámicos, regresión *spline*.

### Abstract

---

<sup>a</sup>Estudiante de doctorado. E-mail: jcsosam@unal.edu.co

<sup>b</sup>Profesor asociado. E-mail: lgdiazm@unal.edu.co

A methodology to estimate time-varying coefficient model's components through generalized estimation equations (Liang & Zeger 1986) is proposed, in order to include directly in the estimation the possible correlation between repeated measurements of each subject. Expansion of the time-varying coefficients is done by means of regression *spline* methods (Huang et al. 2002). Furthermore, is proposed the use of the *Akaike's information criterion in generalized estimating equations* (QIC) proposed by Pan (2001) like model selector. Through simulation are compared the proposed methodology and the methodology presented by Wu & Zhang (2006), where model's components are estimated through weighted least squares and *Akaike's information criterion* (AIC) is used like model selector. It resulted that the proposed methodology gives a better behavior in relation with the average mean square error. In order to illustrate the methodology, is taken into account the data base ACTG 315 (Liang et al. 2003) related to a AIDS study, where it is investigated the relationship between the viral charge and the CD4+ cell count.

**Key words:** AIC, Generalized estimation equations, Regression spline, Longitudinal data, Time-varying coefficient model, Weighted least squares.

## 1. Introducción

El análisis de datos longitudinales surge cuando un conjunto de  $n$  individuos es observado reiteradamente a través del tiempo, registrando los valores de la respuesta de interés junto con las respectivas covariables que pueden depender o no del instante en que sean medidas. Debido a la naturaleza misma de este tipo de datos, una característica fundamental que los distingue, y que debe tenerse en cuenta en el modelamiento, es la posible correlación dada entre las mediciones repetidas de la variable respuesta en cada individuo, considerando las mediciones entre los individuos independientes. Esto es, las mediciones son correlacionadas dentro e independientes entre individuos. Así, se quiere identificar la evolución de la variable respuesta y determinar cómo es afectada por las covariables. Por ejemplo, en estudios de medicina, interesa evaluar el efecto de la dosis de un medicamento u otros factores asociados, como el género del paciente, sobre el progreso de alguna enfermedad en el tiempo.

Las técnicas paramétricas para el análisis de datos longitudinales han sido estudiadas extensivamente en la literatura (Diggle et al. 1994, Davis 2002, Verbeke & Molenberghs 2000, Fitzmaurice et al. 2009). Aunque el enfoque paramétrico es útil, siempre surgirán dudas sobre la adecuación de los supuestos del modelo y el impacto potencial de la falta de especificaciones del modelo sobre el análisis (Hoover et al. 1998).

Las técnicas no paramétricas han sido introducidas recientemente en el análisis de datos longitudinales ya que permiten una dependencia funcional más flexible de la variable respuesta sobre las covariables. Hart & Wehrly (1986), Altman (1990) y Hart (1991) desarrollaron métodos utilizando funciones *kernel* para la estimación de la esperanza de la variable respuesta, sin la presencia de covariables, y propusieron algunas técnicas de selección de parámetros de suavizamiento a través de

la validación cruzada. Estos métodos consideran solo el posible efecto del tiempo sobre la variable respuesta. Para tener en cuenta la influencia de covariables, Zeger & Diggle (1994) estudiaron un modelo semiparamétrico de la forma

$$y_{ij} = \mu(t_{ij}) + \mathbf{x}_{ij}^T(t_{ij})\boldsymbol{\beta} + e_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, n \quad (1)$$

donde  $n$  es el número de individuos en estudio,  $n_i$  es la cantidad de mediciones del  $i$ -ésimo individuo, y  $t_{ij}$ ,  $y_{ij} \equiv y_i(t_{ij})$ ,

$$\mathbf{x}_i(t_{ij}) = [x_{i0}(t_{ij}), x_{i1}(t_{ij}), \dots, x_{id}(t_{ij})]^T$$

y  $e_{ij} \equiv e_i(t_{ij})$  son respectivamente el instante, la variable respuesta de valor real, el vector de covariables en  $\mathbb{R}^{d+1}$  y el error aleatorio, asociados con la  $j$ -ésima medición del  $i$ -ésimo individuo. Además  $\mu(t)$  es una función de  $t$ , y  $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_d]^T$  en  $\mathbb{R}^{d+1}$  es un vector de parámetros. En el contexto de los datos longitudinales  $e_i(t)$  es asumido como un proceso estocástico con media cero y función de covarianza  $\rho_{e_i}(s, t) = Cov(e_i(s), e_i(t))$ .

Hoover et al. (1998) proponen una extensión del modelo (1) donde los parámetros pueden variar con el tiempo. La extensión es de la forma

$$y_{ij} = \mathbf{x}_{ij}^T(t_{ij})\boldsymbol{\beta}(t_{ij}) + e_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, n \quad (2)$$

donde  $\boldsymbol{\beta}(t) = [\beta_{i0}(t_{ij}), \beta_{i1}(t_{ij}), \dots, \beta_{id}(t_{ij})]^T$  es un vector de funciones de  $t$  llamadas coeficientes dinámicos. Este modelo, denominado *modelo de coeficientes dinámicos* (MCD), ha sido estudiado por Wu & Zhang (2006) quienes implementan el método de mínimos cuadrados ponderados junto con la reparametrización de las componentes dinámicas del modelo mediante regresión *spline*. Una de las limitaciones de esta metodología es que no incorpora directamente en el estimador la posible correlación de las mediciones de cada individuo, característica primordial de los datos longitudinales que sí incorpora el estimador propuesto mediante una matriz de correlación de trabajo.

Este artículo está estructurado como sigue. En la sección 2 se revisa brevemente la regresión *spline* y la reparametrización de un MCD utilizando esta técnica. En la sección 3 se revisa la estimación de los parámetros asociados a través del método de mínimos cuadros ponderados y el AIC como selector de modelos. En la sección 4 se propone una alternativa de estimación de los coeficientes mediante las ecuaciones de estimación generalizadas (Liang & Zeger 1986) y el QIC (Pan 2001) como selector de modelos, metodología que no había sido considerada en la estimación de las componentes dinámicas de un MCD. En la sección 5 se muestra un estudio de simulación donde se comparan las metodologías consideradas con base en el error cuadrático medio promedio. En la sección 6 se considera el conjunto de datos ACTG 315 (Liang et al. 2003) asociado con un estudio de sida, en el que se modela dinámicamente la relación entre la carga viral y el conteo de células CD4+. En la sección 7 se discuten los resultados obtenidos y otras alternativas de estimación.

## 2. Regresión *spline*

En esta sección se presenta la estrategia de estimación de las componentes de un MCD utilizando *regresión spline* (RS). Primero se muestran los conceptos básicos involucrados en la RS y en seguida la estrategia de estimación.

### 2.1. Conceptos básicos

**Polinomio a trozos.** Una RS es considerada como un polinomio a trozos. Una función de valor real  $f(\cdot)$  definida sobre el intervalo  $[a, b]$ , es un polinomio a trozos de orden  $k$ , o de grado  $k - 1$ ,  $k \geq 1$ , si es obtenida dividiendo el intervalo  $[a, b]$  en subintervalos contiguos, de tal modo que se pueda representar la función  $f(\cdot)$  mediante un polinomio de orden  $k$  en cada subintervalo. Cada subintervalo es de la forma

$$[\tau_l, \tau_{l+1}), \quad l = 0, \dots, K$$

donde

$$a = \tau_0 < \tau_1 < \dots < \tau_K < \tau_{K+1} = b$$

y  $K + 1$  es el número de subintervalos. Los puntos  $\tau_l$ ,  $l = 0, \dots, K$ , se llaman nodos interiores o simplemente nodos. Note que

$$[a, b) = \bigcup_{l=0}^K [\tau_l, \tau_{l+1}).$$

**Regresión *spline*.** Una RS de orden  $k + 1$ , o de grado  $k$ ,  $k \geq 0$ , con nodos interiores  $\tau_l$ ,  $l = 0, \dots, K$ , es un polinomio a trozos de orden  $k + 1$ , y tiene hasta  $k - 1$  derivadas continuas.

Ya que el espacio de funciones de RS es un espacio vectorial de dimensión finita, hay muchas bases para representar dichas funciones, entre otras, se encuentran: la base de polinomios, la base de potencias truncadas y la base de *B-splines*. Una revisión completa y detallada acerca de *B-splines* se encuentra en de Boor (1978).

**Base de *B-splines*.** Siguiendo a Hastie et al. (1990), antes de definir una *base de B-splines* (BBS), es necesario refinar la secuencia de nodos

$$a = \tau_0 < \tau_1 < \dots < \tau_K < \tau_{K+1} = b$$

Una secuencia de nodos aumentada se define como una secuencia de nodos tal que

- $\xi_1 \leq \xi_2 \leq \dots \leq \xi_M \leq \tau_0$ ;
- $\xi_{j+M} = \tau_j$  para  $j = 1, \dots, K$ ;
- $\tau_{K+1} \leq \xi_{K+M+1} \leq \xi_{K+M+2} \leq \dots \leq \tau_{k+2M}$ .

Los valores de los nodos adicionales más allá de la frontera del intervalo  $[a, b]$  son arbitrarios, y usualmente están dados por

$$\xi_1 = \xi_2 = \dots \xi_M = \tau_0 \quad \text{y} \quad \tau_{K+1} = \xi_{K+M+1} = \xi_{K+M+2} = \dots = \tau_{k+2M}$$

Sea  $B_{i,m}$  la  $i$ -ésima función *B-spline* de orden  $k + 1$ , o grado  $k$ ,  $0 \leq k \leq M - 1$ , asociado a la secuencia de nodos

$$\{\xi_j : j = 1, \dots, k + 2M\},$$

para  $i = 1, \dots, K + 2m - k - 1$ . Los *B-splines* se definen recursivamente como sigue:

$$B_{i,1}(x) = \begin{cases} 1, & \text{si } x \in [\xi_i, \xi_{i+1}] \\ 0, & \text{si } x \notin [\xi_i, \xi_{i+1}] \end{cases} \quad (3)$$

para  $i = 1, \dots, K + 2M - 1$ . Y para  $k \geq 1$  se define

$$B_{i,k+1}(x) = \frac{x - \xi_i}{\xi_{i+k} - \xi_i} B_{i,k}(x) + \frac{\xi_{i+k+1} - x}{\xi_{i+k+1} - \xi_{i+1}} B_{i+1,k}(x) \quad (4)$$

para  $i = 1, \dots, K + 2M - k - 1$ .

Como puede haber algunos nodos duplicados, se debe tener cuidado con las posibles divisiones por cero en (3) y (4). Por lo tanto se adopta la convención  $B_{i,1} = 0$  si  $\xi_i = \xi_{i+1}$ , o si  $\xi_{i+1} = \xi_{i+k+1}$ .

Así, una RS de orden  $m + 1$ , o de grado  $m$ ,  $m \geq 0$ , con nodos interiores  $\tau_r$ ,  $r = 1, \dots, K$ , puede expresarse usando el siguiente conjunto de *splines* base:

$$\mathcal{B} = \{B_{i,m+1} : i = 1, \dots, K + 2M - m - 1\} \quad (5)$$

El conjunto (5) de  $K + 2M - m - 1$  funciones base se conoce como BBS de orden  $m + 1$ , o grado  $m$ , con nodos  $\xi_j$ ,  $j = K + 2M - m - 1$ . Usando las funciones de este conjunto, es posible expresar una RS de orden  $k + 1$  como

$$f(t) = \sum_{i=1}^G \beta_i B_{i,m+1}(t) \quad (6)$$

donde  $\beta_1, \dots, \beta_G$  son los escalares asociados y  $G = K + 2M - m - 1$ . Una revisión completa y detallada acerca de *B-splines* se encuentra en de Boor (1978).

## 2.2. Estimación utilizando regresión *spline*

La idea fundamental de la estrategia de estimación a través de RS es expresar cada componente de  $\beta(t)$  como una RS, escribiendo cada coeficiente dinámico como una combinación lineal de funciones base, de BBS por ejemplo (Huggins & Loesch 1998, Huang et al. 2002).

La idea básica consiste en expresar cada  $\beta_r(t)$ ,  $r = 0, 1, \dots, d$ , como

$$\beta_r(t) = \phi_{r1}(t)\alpha_{r1} + \dots + \phi_{rp_r}(t)\alpha_{rp_r} = \Phi_{rp_r}(t)^T \alpha_r \quad (7)$$

donde  $\Phi_{rp_r}(t) = [\phi_{r1}(t), \dots, \phi_{rp_r}(t)]^T$  es un vector de  $p_r \times 1$  *splines* base, que pueden ser elementos de (5), y  $\alpha_r = [\alpha_{r1}, \dots, \alpha_{rp_r}]^T$  es el vector de  $p_r \times 1$  coeficientes asociados.

Remplazando en el MCD (2) cada  $\beta_r(t)$  por su expresión equivalente empleando funciones base y acomodando los vectores de una forma apropiada se tiene que:

$$\begin{aligned} y_{ij} &= [x_{i0}(t_{ij}), \dots, x_{id}(t_{ij})] \begin{bmatrix} \Phi_{0p_0}(t_{ij})^T \alpha_0 \\ \vdots \\ \Phi_{dp_d}(t_{ij})^T \alpha_d \end{bmatrix} + e_{ij} \\ \Rightarrow y_{ij} &= [x_{i0}(t_{ij})\mathbf{h}_{0ij}^T, \dots, x_{id}(t_{ij})\mathbf{h}_{dij}^T] \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_d \end{bmatrix} + e_{ij} \\ \Rightarrow y_{ij} &= [\mathbf{x}_{0ij}^T, \dots, \mathbf{x}_{dij}^T] \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_d \end{bmatrix} + e_{ij} \\ \Rightarrow y_{ij} &= \mathbf{x}_{ij}^T \alpha + e_{ij} \end{aligned}$$

para cada  $j = 1, \dots, n_i$  y cada  $i = 1, \dots, n$ , donde

$$\mathbf{h}_{rij} = \Phi_{rp_r}(t_{ij}), \quad \mathbf{x}_{ij} = [\mathbf{x}_{0ij}^T, \dots, \mathbf{x}_{dij}^T]^T, \quad \mathbf{x}_{rij} = x_{ri}(t_{ij})\mathbf{h}_{rij}$$

para cada  $r = 0, 1, \dots, d$ , con  $x_{ri}(t)$  la  $r$ -ésima componente de  $\mathbf{x}_i(t)$ , y  $\alpha = [\alpha_0^T, \dots, \alpha_d^T]^T$

Note que si el vector de coeficientes  $\alpha$  de  $p \times 1$ , con  $p = \sum_{r=1}^d p_r$ , está completamente especificado en el modelo equivalente

$$y_{ij} = \mathbf{x}_{ij}^T \alpha + e_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, n \quad (8)$$

entonces cada uno de los parámetros dinámicos del modelo (2) también está determinado. Así, el objetivo es estimar el vector  $\alpha$  a través de algún método apropiado.

El modelo (8) escrito en forma vectorial es

$$\mathbf{y}_i = \mathbf{X}_i^T \alpha + \mathbf{e}_i, \quad i = 1, \dots, n \quad (9)$$

donde

$$\mathbf{y}_i = [y_{i1}, \dots, y_{in_i}]^T$$

es el vector de medidas repetidas,

$$\mathbf{e}_i = [e_{i1}, \dots, e_{in_i}]^T$$

es el vector de errores y

$$\mathbf{X}_i = [\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}]^T$$

es la matriz de diseño, asociados con el  $i$ -ésimo individuo. Además, el modelo (9) puede escribirse en forma matricial como

$$\mathbf{y} = \mathbf{X}\alpha + \mathbf{e}, \quad (10)$$

donde  $\mathbf{y} = [\mathbf{y}_1^T, \dots, \mathbf{y}_n^T]^T$ ,  $\mathbf{X} = [\mathbf{X}_1^T, \dots, \mathbf{X}_n^T]^T$  y  $\mathbf{e} = [\mathbf{e}_1^T, \dots, \mathbf{e}_n^T]^T$ .

En este punto, debido a la expresión de las componentes dinámicas del modelo (2) a través de *B-splines*, este es equivalente a un modelo de regresión lineal en el que se requiere estimar el vector de parámetros  $\boldsymbol{\alpha}$ .

Una vez obtenida una estimación de los parámetros del modelo (10), denotada por  $\hat{\boldsymbol{\alpha}} = [\hat{\boldsymbol{\alpha}}_0^T, \dots, \hat{\boldsymbol{\alpha}}_d^T]^T$ , la estimación de los coeficientes  $\beta_r(t)$  está dada por

$$\hat{\beta}_r(t) = \boldsymbol{\Phi}_{rp_r}(t)^T \hat{\boldsymbol{\alpha}}_r, \quad r = 0, 1, \dots, d \quad (11)$$

donde  $\boldsymbol{\Phi}_{rp_r}(t)$  es el vector de *splines* base asociados con el  $r$ -ésimo parámetro dinámico.

La elección de los parámetros de suavizamiento está relacionada directamente con el conjunto de *splines* base con que se expresen los coeficientes dinámicos del MCD (2). Por ejemplo, trabajando con BBS, están dados por

$$p_r = K_r - k_r - 1, \quad r = 0, 1, \dots, d \quad (12)$$

donde  $K_r$  es el número de nodos y  $k_r$  es el grado de la base asociada con la estimación del  $r$ -ésimo coeficiente dinámico. Entonces, seleccionar el valor de los parámetros de suavizamiento es equivalente a seleccionar los  $K_r$  y los  $k_r$ . Para ello, es costumbre fijar el grado de la base  $k_r$ , por ejemplo 1, 2 ó 3, es decir, lineal, cuadrático o cúbico, y en seguida elegir mediante un criterio adecuado el número de nodos  $K_r$  para determinar el valor del parámetro de suavizamiento  $p_r$ .

Cuando se usa una BBS para expresar los parámetros dinámicos del MCD (2), una vez fijado el número de nodos  $K_r$  de cada base a través de algún criterio apropiado, se deben ubicar en el rango de interés, que en el caso de los datos longitudinales es sobre el conjunto de los tiempos de medición

$$\{t_{ij} : i = 1, \dots, n; j = 1, \dots, n_i\}$$

En cuanto a la ubicación de los nodos se distinguen dos métodos:

**Método 1.** Consiste en ubicar los  $K_r$  nodos igualmente espaciados en el intervalo de interés  $[a, b]$ . En el caso de los datos longitudinales, los valores que definen este intervalo son el mínimo y el máximo de todos los tiempos, esto es

$$a = \min\{t_{ij} : i = 1, \dots, n; j = 1, \dots, n_i\}$$

$$b = \max\{t_{ij} : i = 1, \dots, n; j = 1, \dots, n_i\}$$

**Método 2.** Para ubicar los  $K_r$  nodos, según este método, se deben considerar los  $M$  tiempos diferentes

$$t_1 < t_2 < \dots < t_M \quad (13)$$

El método consiste en ubicar los nodos igualmente espaciados sobre los cuantiles de los tiempos (13).

### 3. Estimación a través del método de mínimos cuadrados ponderados

Una alternativa de estimación de los parámetros del modelo (10) es a través del método de *mínimos cuadrados ponderados* (MCP). Este método consiste en estimar  $\alpha$  minimizando la suma de cuadrados

$$\sum_{i=1}^n \sum_{j=1}^{n_i} w_i (y_{ij} - \mathbf{x}_{ij}^T \alpha)^2 \quad (14)$$

donde los  $w_i$  son pesos dados usando alguno de los siguientes esquemas de ponderación:

**Esquema 1.** En el que los pesos están dados por  $w_i = 1/N$ ,  $i = 1, \dots, n$ , donde  $N = \sum_{i=1}^n n_i$ .

**Esquema 2.** En el que los pesos están dados por  $w_i = 1/(nn_i)$ ,  $i = 1, \dots, n$ .

El esquema 1 usa el mismo peso para todos los individuos y fue implementado por Hoover et al. (1998). El esquema 2 es considerado por Huang et al. (2002) y usa diferentes pesos para los individuos. Huang et al. (2002) demuestran que el esquema 1 puede llevar a estimaciones inconsistentes de  $\alpha$ .

Minimizando la función objetivo (14) trabajando con el modelo (10), se obtiene un estimador de  $\alpha$  dado por

$$\hat{\alpha}_{MCP} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y} \quad (15)$$

donde  $\mathbf{W} = \text{diag}[\mathbf{W}_1, \dots, \mathbf{W}_n]$ , con  $\mathbf{W}_i = w_i \mathbf{I}_{n_i}$  la matriz de pesos del  $i$ -ésimo individuo,  $i = 1, \dots, n$ , e  $\mathbf{I}_{n_i}$  la matriz identidad de  $n_i \times n_i$ .

**Criterio de información de Akaike (AIC).** La idea básica del AIC, concebido para la estrategia de estimación que involucra el método de MCP, es encontrar la combinación de parámetros de suavizamiento, determinados por la cantidad de nodos, que minimicen la expresión

$$AIC(\rho) = -2\text{Loglik} + 2df, \quad (16)$$

donde  $\rho = [p_0, p_1, \dots, p_d]^T$  es el vector conformado por los parámetros de suavizamiento,

$$\text{Loglik} = -\frac{n}{2} \log \left( \frac{2\pi e}{n} SCE_\rho \right) \quad (17)$$

con

$$SCE_\rho = \sum_{i=1}^n \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2$$

y

$$df = \text{tr}(\mathbf{A}) \quad (18)$$



donde  $\mathbf{A}$  es la matriz de suavizamiento asociada con el estimador, que en el caso del estimador (15) está dada por

$$\mathbf{A} = \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}$$

Se elige el modelo que minimice la cantidad  $AIC(\boldsymbol{\rho})$ . Esta medida permite encontrar un equilibrio entre la bondad de ajuste del modelo, representada por  $Loglik$ , y la complejidad del modelo, representada por  $df$ . Es decir, la bondad de ajuste del modelo es penalizada con la complejidad del mismo (Wu & Zhang 2006).

Una de las desventajas de este método es que no tiene en cuenta directamente la posible correlación de las medidas repetidas de cada individuo, una de las características más importantes de la estructura de los datos longitudinales. En la siguiente sección se propone una alternativa de estimación que considera la estructura de correlación de las medidas repetidas.

## 4. Estimación a través de las ecuaciones de estimación generalizadas

Aquí se propone otra alternativa de estimación de los parámetros del modelo (10) empleando las *ecuaciones de estimación generalizadas* (EEG) de Liang & Zeger (1986), en las que se asume una matriz de correlación de trabajo específica para la componente del error, con el fin de incluir directamente la posible correlación de las medidas repetidas de cada individuo. Una de las ventajas de este método es que, sin importar la matriz de correlación de trabajo seleccionada, las EEG siempre llevan a estimaciones consistentes y asintóticamente normales, aunque elegir una estructura de correlación adecuada incrementa la eficiencia del método (Davis 2002, pág. 296).

### 4.1. Metodología

El método de estimación empleando las EEG requiere la selección de una función de enlace, una función de varianza y una matriz de correlación de trabajo. Una vez seleccionadas estas componentes, según lo indique la naturaleza de los datos, la estimación de los parámetros del modelo (10), denotada por  $\hat{\boldsymbol{\alpha}}_{EEG}$ , corresponde a la solución para  $\boldsymbol{\alpha}$  del sistema de ecuaciones

$$\sum_{i=1}^n \left( \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\alpha}} \right)^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) = \mathbf{0} \quad (19)$$

donde  $\boldsymbol{\mu}_i = [\boldsymbol{\mu}_{i1}, \dots, \boldsymbol{\mu}_{in_i}]^T$  es el vector de medias asociado con las medidas repetidas del  $i$ -ésimo individuo, esto es,  $\mu_{ij} = \mathbb{E}(y_{ij})$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, n_i$ ,  $\hat{\boldsymbol{\theta}}$  es un estimador consistente de  $\boldsymbol{\theta}$ , el vector de parámetros asociado con la matriz de correlación de trabajo del  $i$ -ésimo individuo  $\mathbf{R}_i \equiv \mathbf{R}_i(\boldsymbol{\theta})$ ,

$$\mathbf{V}_i \equiv \mathbf{V}_i(\boldsymbol{\theta}) = \phi \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\theta}) \mathbf{A}_i^{1/2}, \quad i = 1 \dots, n \quad (20)$$

es la matriz de covarianzas de trabajo, con  $\phi$  un parámetro de escala posiblemente desconocido, y

$$\mathbf{A}_i = \text{diag}[V(\mu_{i1}), \dots, V(\mu_{in_i})], \quad i = 1, \dots, n$$

con  $V(\cdot)$  la función de varianza correspondiente.

En Davis (2002, pág. 300) se presenta el proceso correspondiente para resolver el sistema de ecuaciones (19) y también un procedimiento para hallar una estimación consistente de la varianza estimada de  $\hat{\boldsymbol{\alpha}}_{EEG}$ .

**Criterio de información de Akaike en las ecuaciones de estimación generalizadas (QIC).** Considere el conjunto de datos

$$\mathcal{D} = \{(y_{ij}, \mathbf{x}_{ij}) : i = 1, \dots, n; j = 1, \dots, n_i\}$$

donde  $y_{ij}$  y  $\mathbf{x}_{ij}$  son el valor de la variable respuesta y el vector de covariables asociados con la  $j$ -ésima medición del  $i$ -ésimo individuo, y  $\boldsymbol{\alpha}$  es el vector de parámetros, que puede estimarse a través de las EEG usando cualquier estructura de correlación de trabajo.

Siguiendo a Pan (2001), bajo el modelo de independencia, la cuasiverosimilitud conjunta, basada en el conjunto de datos  $\mathcal{D}$ , está dada por

$$Q(\boldsymbol{\alpha}, \phi; \mathbf{I}, \mathcal{D}) = \sum_{i=1}^n \sum_{j=1}^{n_i} Q(\boldsymbol{\alpha}, \phi; y_{ij}, \mathbf{x}_{ij}) \quad (21)$$

donde  $\phi$  es un parámetro de escala posiblemente desconocido y

$$Q(\boldsymbol{\alpha}, \phi; y, \mathbf{x}) = Q(g^{-1}(\mathbf{x}^T \boldsymbol{\alpha}), \phi; y) = \int_y^\mu \frac{y-t}{\phi V(t)} dt \quad (22)$$

con  $\mathbb{V}(y) = \phi V(\mu)$ ,  $V(\cdot)$  una función de varianza,  $\mu = \mathbb{E}(y) = g^{-1}(\mathbf{x}^T \boldsymbol{\alpha})$ ,  $g(\cdot)$  una función de enlace, y  $\mathbf{x}$  un vector de covariables.

La idea básica del criterio, concebido para la estimación utilizando las EEG, es minimizar una medida basada en  $Q(\hat{\boldsymbol{\alpha}}, \hat{\phi}; \mathbf{I}, \mathcal{D})$ , la cuasiverosimilitud bajo el modelo de trabajo de independencia con una estimación de  $\boldsymbol{\alpha}$ , usando cualquier matriz de correlación de trabajo empleada en las EEG.

El modelo de independencia es asumido para la construcción de la cuasiverosimilitud conjunta, aunque en las EEG se puede emplear cualquier estructura de correlación. Este criterio se puede generalizar utilizando cualquier matriz de correlación en la construcción de la cuasiverosimilitud conjunta, pero puede que esta no sea única (Pan 2001).

El QIC consiste en encontrar la combinación de parámetros de suavizamiento, determinados por la cantidad de nodos, que minimicen la expresión

$$QIC = -2Q(\hat{\boldsymbol{\alpha}}, \hat{\phi}; \mathbf{I}, \mathcal{D}) + 2tr(\hat{\boldsymbol{\Omega}}, \hat{\mathbf{V}}_r) \quad (23)$$

donde  $\hat{\alpha} \equiv \hat{\alpha}(\mathbf{R})$  es un estimador del vector de parámetros  $\alpha$  del modelo (10), obtenido a través del método de las EEG usando cualquier estructura de correlación  $\mathbf{R}$ ,  $\hat{\mathbf{V}}_r$  es cualquier estimación consistente de  $\mathbb{V}(\hat{\alpha})$ , como el estimador del *sandwich* (Liang & Zeger 1986) y

$$\hat{\Omega} = -\partial^2 Q(\alpha, \hat{\phi}; \mathbf{I}, \mathcal{D}) / \partial \alpha \partial \alpha^T |_{\alpha = \hat{\alpha}} d$$

### 4.2. Propiedades

Sea  $\hat{\beta}_{EEG}(t)$  el estimador de  $\beta(t) = [\beta_0(t), \beta_1(t), \dots, \beta_d(t)]^T$  utilizando la metodología propuesta apoyada en las EEG. Este estimador está dado por

$$\hat{\beta}_{EEG}(t) = [\hat{\beta}_0(t), \hat{\beta}_1(t), \dots, \hat{\beta}_d(t)]^T = \Phi(t)^T \hat{\alpha}_{EEG}$$

donde  $\Phi(t)$  es la matriz de  $\sum_{r=0}^d p_r \times (d + 1)$  dada por

$$\Phi \equiv \Phi(t) = \begin{bmatrix} \Phi_{0p_0}(t) & \mathbf{0}_{p_0} & \cdots & \mathbf{0}_{p_0} \\ \mathbf{0}_{p_1} & \Phi_{1p_1}(t) & \vdots & \mathbf{0}_{p_1} \\ \vdots & \cdots & \ddots & \vdots \\ \mathbf{0}_{p_d} & \mathbf{0}_{p_d} & \cdots & \Phi_{dp_d}(t) \end{bmatrix}$$

con  $\mathbf{0}_{p_r}$  un vector columna constituido por  $p_r$  ceros,  $r = 0, 1, \dots, d$ .

En un tiempo dado, como  $\hat{\beta}_{EEG}(t)$  es una transformación lineal de  $\hat{\alpha}_{EEG}$ , el siguiente corolario derivado del teorema 2 de Liang & Zeger (1986, pág. 16) establece la distribución asintótica de  $\hat{\beta}_{EEG}(t)$ .

**Corolario 1.** *Bajo condiciones de regularidad y dado que:*

- (i)  $\hat{\theta}$  es un estimador  $\sqrt{n}$ -consistente de  $\theta$  dados  $\alpha$  y  $\phi$ ;
- (ii)  $\hat{\phi}$  es un estimador  $\sqrt{n}$ -consistente de  $\phi$  dado  $\alpha$ ; y
- (iii)  $\|\partial \hat{\theta}(\alpha, \phi) / \partial \phi\| = O_p(1)$ ,

entonces asintóticamente  $\sqrt{n}(\hat{\beta}_{EEG}(t) - \beta(t))$  tiene una distribución normal multivariada con media cero y matriz de covarianza

$$\lim_{n \rightarrow \infty} n \Phi \mathbf{M}_0^{-1} \mathbf{M}_1 \mathbf{M}_0^{-1} \Phi^T$$

donde

$$\mathbf{M}_0 = \sum_{i=1}^n \left( \frac{\partial \mu_i}{\partial \alpha} \right)^T \mathbf{V}_i^{-1} \left( \frac{\partial \mu_i}{\partial \alpha} \right) \tag{24}$$

y

$$\mathbf{M}_1 = \sum_{i=1}^n \left( \frac{\partial \mu_i}{\partial \alpha} \right)^T \mathbf{V}_i^{-1} \mathbb{V}(\mathbf{y}_i) \mathbf{V}_i^{-1} \left( \frac{\partial \mu_i}{\partial \alpha} \right) \tag{25}$$

Así, un estimador de la varianza de  $\widehat{\beta}_{EEG}(t)$  está dado por

$$\widehat{\mathbb{V}}(\widehat{\beta}_{EEG}(t)) = \Phi \widehat{\mathbf{M}}_0^{-1} \widehat{\mathbf{M}}_1 \widehat{\mathbf{M}}_0^{-1} \Phi^T$$

donde  $\widehat{\mathbf{M}}_0$  y  $\widehat{\mathbf{M}}_1$  son obtenidos como en las expresiones (24) y (25) reemplazando  $\mathbb{V}(\mathbf{y}_i)$  por  $(\mathbf{y}_i - \widehat{\boldsymbol{\mu}})(\mathbf{y}_i - \widehat{\boldsymbol{\mu}})^T$  y  $\boldsymbol{\alpha}$ ,  $\phi$  y  $\boldsymbol{\theta}$  por sus respectivos estimadores. Este estimador de la varianza de  $\widehat{\beta}_{EEG}(t)$  es un estimador consistente de  $\mathbb{V}(\widehat{\beta}_{EEG}(t))$  aunque la matriz de correlación de trabajo  $\mathbf{R}_i$  no corresponda a la verdadera matriz de correlación de  $\mathbf{y}_i$  (ver la prueba en el apéndice A).

## 5. Simulación

Con el fin de evaluar el desempeño de los métodos de estimación, la comparación entre ellos se hace a través del *error cuadrático medio promedio* (ECMP) dado por

$$ECMP(\kappa) = \frac{1}{n} \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} [\kappa(t_{ij}) - \widehat{\kappa}(t_{ij})]^2 \quad (26)$$

donde  $\kappa(\cdot)$  es una función que corresponde a cualquier coeficiente dinámico del modelo.

La estrategia de simulación utilizada es similar a la presentada en Wu & Liang (2004). El modelo empleado en la simulación es de la forma

$$y_i(t) = \beta_0(t) + x_{i1}(t)\beta_1(t) + e_i(t) \quad (27)$$

donde  $\beta_0(t)$  y  $\beta_1(t)$  son los coeficientes dinámicos del modelo,  $x_i(t)$  es la única covariable del modelo, asociada con  $\beta_1(t)$ , y  $e_i(t)$  es el error en el tiempo  $t$ . Note que este modelo corresponde a un MCD dado por (2) donde el vector de parámetros dinámicos es  $\boldsymbol{\beta}(t) = [\beta_0(t), \beta_1(t)]^T$  y el vector de covariables es  $\mathbf{x}_i(t) = [x_{i0}(t), x_{i1}(t)]^T$  con  $x_{i0}(t) \equiv 1$ .

En el MCD simulado la covariable  $x_{i1}(t)$  está dada por

$$x_{i1}(t) = 1 - \exp\left(-0.5t - \frac{i}{n}\right), \quad i = 1, \dots, n \quad (28)$$

y los parámetros dinámicos se definen como

$$\beta_0(t) = 3 \exp(t), \quad \text{y} \quad \beta_1(t) = 1 + \cos(2\pi t) + \sin(2\pi t) \quad (29)$$

En la figura 1 se muestran los coeficientes dinámicos simulados.

El término de error correspondiente al modelo (27) se simula bajo dos estructuras de correlación:

**Estructura de correlación 1.** En la que se considera a  $e_i(t)$  distribuido normalmente con media 0 y varianza  $\sigma_e^2 x_{i1}^2(t)$ , donde las medidas repetidas entre individuos y dentro de cada individuo son independientes.

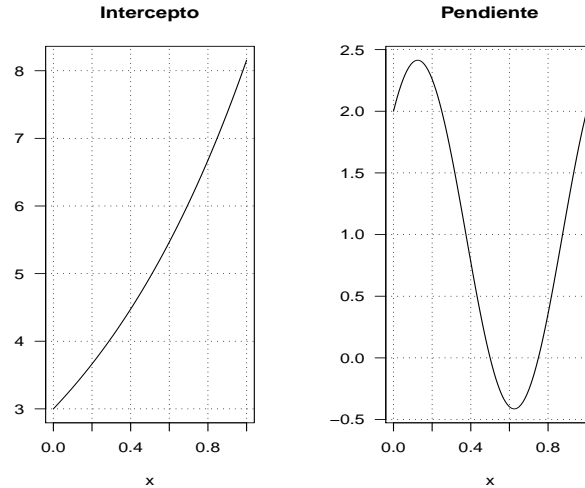


FIGURA 1: Coeficientes dinámicos simulados.

**Estructura de correlación 2.** En la que se considera al vector de errores asociados con las medidas repetidas de un individuo distribuido normalmente con vector de medias  $\mathbf{0}$  y matriz de covarianzas asociada con una estructura de correlación autoregresiva de primer orden, esto es, una matriz de correlación  $\mathbf{R} = [R_{kl}]$  donde

$$R_{kl} = \begin{cases} \alpha^{|k-l|}, & \text{si } k \neq l \\ 1, & \text{si } k = l \end{cases}$$

con  $0 < \alpha < 1$ .

Los tiempos simulados son de la forma

$$t_{ij} = j/(m+1), \quad i = 1, \dots, n, \quad j = 1, \dots, m$$

siendo  $m$  un entero positivo, y para simular conjuntos de datos no balanceados<sup>1</sup>, una característica propia de la estructura de los datos longitudinales, en cada individuo se remueven aleatoriamente medidas repetidas con una tasa  $r_m$ . Así, en promedio hay  $m(1 - r_m)$  medidas repetidas por cada individuo y  $nm(1 - r_m)$  medidas en total.

Los parámetros de suavizamiento de los métodos respectivos se eligen por medio del AIC y el QIC.

El estudio de simulación se hace en R considerando seis escenarios:

**Escenario 1**  $n = 25$ ,  $m = 8$ ,  $r_m = 20\%$ , y  $\sigma_e^2 = 0.01$  en la estructura de correlación 1.

<sup>1</sup>El número de medidas repetidas varía de individuo a individuo.

**Escenario 2**  $n = 25, m = 8, r_m = 20\%$ , y  $\sigma_e^2 = 0.04$  en la estructura de correlación 1.

**Escenario 3**  $n = 25, m = 8, r_m = 20\%$ , y  $\sigma_e^2 = 0.09$  en la estructura de correlación 1.

**Escenario 4**  $n = 25, m = 8, r_m = 20\%$ , y  $\phi = 1$  y  $\alpha = 0.335$  en la estructura de correlación 2.

**Escenario 5**  $n = 25, m = 8, r_m = 20\%$ , y  $\phi = 1$  y  $\alpha = 0.665$  en la estructura de correlación 2.

**Escenario 6**  $n = 25, m = 8, r_m = 20\%$ , y  $\phi = 1$  y  $\alpha = 0.820$  en la estructura de correlación 2.

Cada escenario se replicó  $N = 200$  veces y en cada ocasión se calculó  $ECMP(\beta_0)$  y  $ECMP(\beta_1)$ , con el fin de comparar el desempeño relativo de la *estimación a través de las ecuaciones de estimación generalizadas* (EEEG) con la *estimación a través de mínimos cuadrados ponderados* (EMCP). Para ello se definen los indicadores

$$ECMPR = \frac{1}{N} \sum_{k=1}^N \frac{ECMP_k(\kappa, EMCP)}{ECMP_k(\kappa, EEEG)} \times 100\% \tag{30}$$

y

$$ECMPG = \frac{\sum_{k=1}^N I_{\{ECMP_k(\kappa, EMCP) > ECMP_k(\kappa, EEEG)\}}}{N} \times 100\% \tag{31}$$

donde  $ECMP_k(\kappa, EMCP)$  y  $ECMP_k(\kappa, EEEG)$  denotan el valor de  $ECMP(\kappa)$  obtenido en la  $k$ -ésima réplica de la simulación,  $k = 1, \dots, N$ , usando EMCP y EEEG respectivamente, y  $I_A$  denota la función indicadora del conjunto  $A$ .

TABLA 1: Resultados de la simulación.

Escenario	$ECMPR_0$	$ECMPR_1$	$ECMPG_0$	$ECMPG_1$
Escenario 1	451.6 %	1358.7 %	100.0 %	100.0 %
Escenario 2	388.0 %	979.5 %	100.0 %	100.0 %
Escenario 3	285.1 %	542.1 %	99.0 %	100.0 %
Escenario 4	97.6 %	165.4 %	38.5 %	43.0 %
Escenario 5	170.3 %	401.7 %	62.0 %	64.0 %
Escenario 6	180.8 %	483.4 %	58.5 %	69.5 %

El  $ECMPR$  representa la eficiencia relativa promedio asociada con las  $N$  réplicas ( $N = 200$  en este caso), y  $ECMPG$  es el porcentaje de estimadores obtenidos a través de EEG, que son mejores que los obtenidos a través de las MCP en cuanto al ECMP en las  $N$  réplicas.

Note que: si  $ECMPR \approx 1$  y  $ECMPG = 50\%$ , la EEEG y la EMCP se desempeñan similarmente; si  $ECMPR > 1$  y  $ECMPG > 50\%$ , la EEEG se desempeña mejor que la EMCP; y si  $ECMPR < 1$  y  $ECMPG < 50\%$ , la EEEG se desempeña peor que la EMCP.

En la tabla 1 se presentan los resultados obtenidos en la simulación. En todos los escenarios considerados, menos el número 4, se obtuvo que  $ECMPR > 1$  y

$ECMPG > 50\%$ , por lo que la EEEG se desempeña mejor que la EMCP con relación al error cuadrático medio promedio. En el escenario 4, como  $ECMPR \approx 1$  y  $ECMPG \approx 50\%$ , se tiene que la EEEG y la EMCP se desempeñan similarmente. Esta similitud se debe a que el parámetro de correlación entre las medidas repetidas  $\alpha$  no es suficientemente grande para marcar diferencias entre la estructura de correlación autorregresiva de primer orden y la estructura de correlación de independencia. Por ejemplo, note que en el escenario 4 la correlación existente entre la primera medida repetida y la tercera es de apenas  $R_{13} = 0.037$ .

## 6. Aplicación

En esta sección se presentan los resultados del análisis del conjunto de datos ACTG 315 (Liang et al. 2003). Este conjunto de datos corresponde a un estudio del *síndrome de inmunodeficiencia adquirida* (sida), llevado a cabo para investigar la relación entre carga viral y número de células CD4+ en individuos infectados con el *virus de inmunodeficiencia humana* (VIH).

### 6.1. Introducción

La carga viral (*plasma VIH RNA copies/mL*) y el conteo de células CD4+ son actualmente indicadores decisivos para evaluar tratamientos de sida en investigación clínica. Inicialmente se consideró el conteo de células CD4+ como un indicador inmunológico primario de sida, pero últimamente se ha encontrado que la carga viral es más predictiva en los resultados clínicos. Sin embargo, recientemente algunos investigadores han sugerido que la combinación de estos dos indicadores puede ser más apropiada para evaluar los tratamientos de VIH y sida. Por ello es pertinente estudiar la relación entre la carga viral y el conteo de células CD4+ durante el tratamiento (Liang et al. 2003).

En la figura 2 se presentan algunos gráficos de una regresión lineal de la carga viral ( $\log_{10}(\text{RNA})$ ) frente al conteo de células CD4+ en algunos tiempos de un estudio clínico de sida (ACTG 315). En este, se trata de 46 pacientes infectados, con una terapia antiviral que consistía de *ritonavir*, 3TC y AZT. Una vez iniciado el tratamiento, la carga viral y el conteo de células CD4+ fueron observados simultáneamente los días 0, 2, 7, 10, 14, 28, 56, 84, 168 y 336. En total se obtuvieron 361 observaciones, y el número de medidas repetidas por individuo varía de 4 a 10.

En general, se observa que la respuesta virológica (medida por la carga viral) y la respuesta inmunológica (medida por el conteo de células CD4+) del paciente están correlacionadas negativamente, y que su relación es aproximadamente lineal durante el tratamiento antiviral. En la figura 3 se presentan los dispersogramas del conteo de células CD4+ y de la carga viral. Se utiliza el logaritmo de la carga viral para estabilizar la varianza en los procedimientos de estimación de los modelos ajustados en las secciones siguientes.

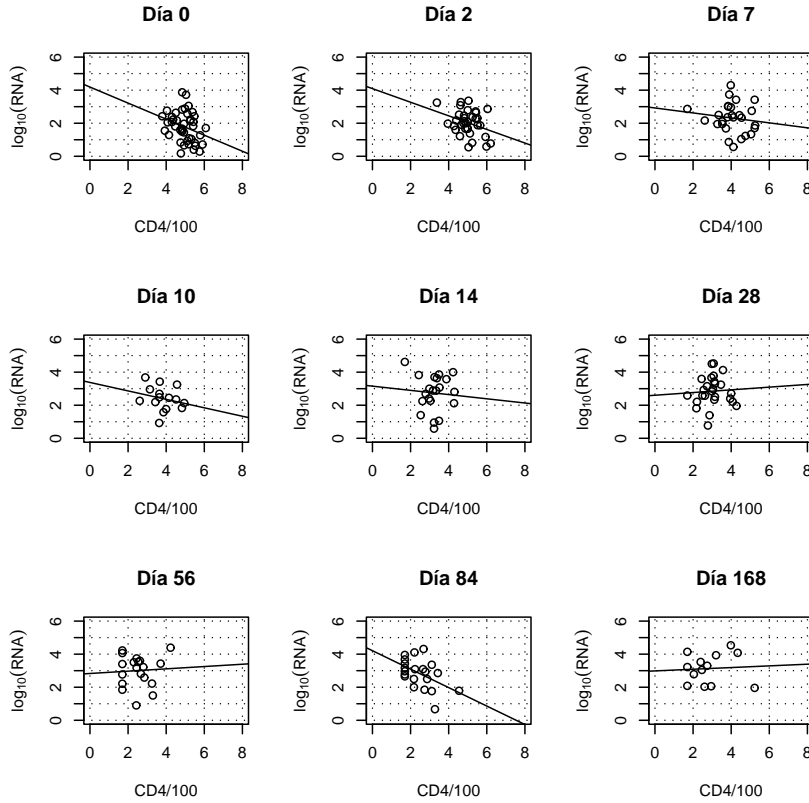


FIGURA 2: Gráficos de una regresión lineal de la carga viral ( $\log_{10}(\text{RNA})$ ) frente al conteo de células CD4+ en algunos tiempos. El modelo ajustado en cada caso es de la forma  $\log_{10}(\text{RNA}) = \beta_0 + \beta_1(\text{CD4}/100) + e$ .

En la figura 2, note que la pendiente de cada regresión lineal cambia con el tiempo. Esto motiva a modelar la carga viral con un MCD.

El conjunto de datos ACTG 315 ya ha sido estudiado ampliamente por Liang et al. (2003), donde se evidencia principalmente una fuerte relación inversa entre la carga viral y el conteo de células CD4+.

## 6.2. Modelamiento

En esta sección se presentan los resultados del análisis del conjunto de datos ACTG 315 usando un MCD. Se ajusta un MCD para investigar la relación dinámica entre la carga viral (en escala logarítmica) y el conteo de células CD4+.

El MCD ajustado es de la forma

$$y_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij})x_{i1}(t_{ij}) + e_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, 46 \quad (32)$$



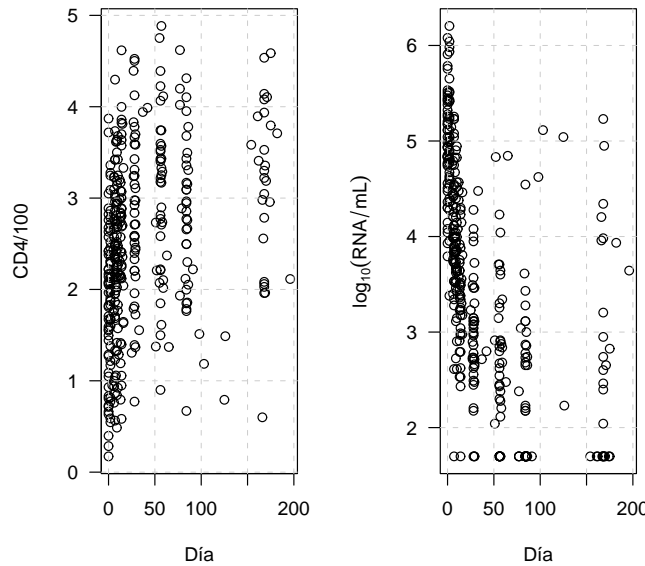


FIGURA 3: Dispersograma del conteo de células CD4+ y de la carga viral.

donde  $y_{ij}$ ,  $x_{i1}(t_{ij})$ , y  $e_{ij}$  son la carga viral (en escala logarítmica), el conteo de células CD4+ y el error asociados con la  $j$ -ésima medición del  $i$ -ésimo individuo, respectivamente, y  $\beta_0(t)$  y  $\beta_1(t)$  son los coeficientes dinámicos del modelo. Note que  $\beta_1(t)$  es el coeficiente relacionado con la relación dinámica entre la carga viral y el conteo de células CD4+.

Las componentes dinámicas del modelo se estiman a través de la metodología propuesta utilizando RS y las EEG. La base de *splines* empleada en la RS es una BBS de segundo grado, y en la selección del número de los nodos se emplea el QIC. La ubicación de los nodos se hace según el método 1 en el que se ubican los nodos igualmente espaciados en el intervalo de interés.

TABLA 2: Parámetros de suavizamiento que minimizan el criterio de selección bajo una estructura de correlación de trabajo dada.

$p_0$	$p_1$	Estructura	$ECMP(\beta_0)$	$ECMP(\beta_1)$
11	10	Independencia	7.450128e+166	8.706318e-02
4	4	Intercambiable	7.450128e+166	8.701916e-02
8	7	Autorregresiva (1)	7.450128e+166	8.698554e-02

Como se muestra en la tabla 2, la estructura de correlación de trabajo que minimiza el ECMP es la estructura intercambiable. Note que el ECMP de todas las estructuras es similar, así que por simplicidad se eligió la estructura de independencia, donde los parámetros de suavizamiento asociados con  $\beta_0(t)$  y  $\beta_1(t)$  que

minimizan el QIC son  $p_0 = 11$  y  $p_1 = 10$ , respectivamente. En este caso, la estructura de correlación no tiene mayor incidencia en la estimación de las componentes del modelo, porque la cantidad de medidas repetidas de cada individuo es pequeña comparada con la cantidad de mediciones total.

En la figura 4 se muestra la estimación del coeficiente dinámico asociado con la pendiente del MCD ajustado. Note que al principio del tratamiento la relación entre la carga viral y el conteo de células CD4+ no es tan fuerte hasta aproximadamente el día 90 del tratamiento donde se evidencia una relación directa hasta el día 100. Entre los días 100 y 140 se nota una relación claramente inversa. Entre los días 140 y 150 parece que la relación es directa, pero se nota que vuelve a ser débil. Estos resultados son similares a los obtenidos por Liang et al. (2003).

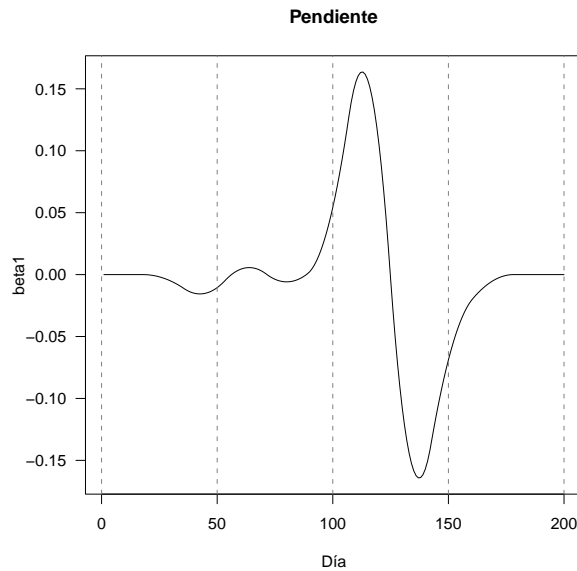


FIGURA 4: Estimación del coeficiente dinámico asociado con la pendiente del MCD ajustado.

## 7. Discusión

La metodología propuesta permite involucrar en la estimación de los coeficientes dinámicos y en la selección de los parámetros de suavizamiento la posible correlación de las medidas repetidas de cada individuo en estudio. Por ello en los escenarios simulados donde el grado de correlación es significativo se obtuvo que utilizar las EEG y el QIC da mejores resultados con relación al error cuadrático medio que utilizar el método de MCP y el AIC.

Incluir en el estimador de las componentes del MCD la posible correlación entre las medidas repetidas no tiene un costo computacional importante, puesto que los

procesos computacionales se pueden implementar fácil y eficientemente utilizando software como la función `geese.fit` del paquete `geepack` de R.

En la metodología propuesta se aproximan los coeficientes dinámicos mediante regresión *spline*, pero cabe notar que no es la única alternativa. También es posible llevar a cabo la aproximación mediante suavizamiento *spline*, polinomios locales *kernel* (Wu & Zhang 2006) o mediante funciones radiales *kernel* (Sosa & Díaz 2009), técnicas a comparar en futuros estudios de simulación.

Utilizando regresión *spline*, también es posible emplear bases diferentes de *B-splines* en la aproximación de los coeficientes dinámicos, por ejemplo bases de potencias truncadas (Ramsay & Silverman 1997). Esta alternativa también es susceptible de investigación en estudios de simulación posteriores.

[Recibido: agosto de 2009 — Aceptado: abril de 2010]

## Referencias

- Altman, N. S. (1990), 'Kernel Smoothing of Data with Correlated Errors', *Journal of the American Statistical Association* **85**(411), 749–759.
- Davis, C. S. (2002), *Statistical Methods for the Analysis of Repeated Measurements*, Springer.
- de Boor, C. (1978), *A Practical Guide to Splines*, Springer.
- Diggle, P. J., Liang, K. Y. & Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford University Press.
- Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. (2009), *Longitudinal Data Analysis*, Chapman & Hall.
- Hart, J. D. (1991), 'Kernel Regression Estimation with Time Series Errors', *Journal of the Royal Statistical Society. Series B (Methodological)* **53**(1), 173–187.
- Hart, J. D. & Wehrly, T. E. (1986), 'Kernel Regression Estimation using Repeated Measurements Data', *Journal of the American Statistical Association* **81**(396), 1080–1088.
- Hastie, T., Tibshirani, R. & Friedman, J. (1990), *The Elements of Statistical Learning*, Springer.
- Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L. P. (1998), 'Nonparametric Smoothing Estimates of Time-Varying Coefficient Models with Longitudinal Data', *Biometrika* **85**(4), 809–822.
- Huang, J. Z., Wu, C. O. & Zhou, L. (2002), 'Varying-coefficient Models and Basis Function Approximations for the Analysis of Repeated Measurements', *Biometrika* **89**(1), 111–128.

- Huggins, R. M. & Loesch, D. Z. (1998), 'On the Analysis of Mixed Longitudinal Growth Data', *Biometrics* **54**(2), 583–595.
- Liang, H., Wu, H. & Carroll, R. J. (2003), 'The relationship between Virologic and Immunologic Responses in AIDS Clinical Research using Mixed-Effects Varying-Coefficient Models with Measurement Error', *Biostatistics* **4**(2), 297–312.
- Liang, K. Y. & Zeger, S. L. (1986), 'Longitudinal Data Analysis using Generalized Linear Models', *Biometrika* **73**(1), 13–22.
- Pan, W. (2001), 'Akaike's Information Criterion in Generalized Estimating Equations', *Biometrics* **57**(1), 120–125.
- Ramsay, J. O. & Silverman, B. W. (1997), *Applied Functional Data Analysis*, Springer.
- Sosa, J. C. & Díaz, L. G. (2009), Desarrollo de un modelo de coeficientes dinámicos y aleatorios para el análisis longitudinales, Tesis de maestría, Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia.
- Verbeke, G. & Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, Springer.
- Wu, H. & Liang, H. (2004), 'Backing Random Varying-Coefficient Models with Time-Dependent Smoothing Covariates', *Scandinavian Journal of Statistics* **31**, 3–19.
- Wu, H. & Zhang, J. T. (2006), *Nonparametric Regression Methods for Longitudinal Data Analysis*, Wiley.
- Zeger, S. L. & Diggle, P. J. (1994), 'Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters', *Biometrics* **50**(3), 689–699.

## Apéndice

Sea  $\widehat{V}_{EEG}$  un estimador consistente de la varianza de  $\widehat{\alpha}_{EEG}$ , como el estimador del *sandwich*, que es un estimador consistente aunque la matriz de correlación de trabajo  $\mathbf{R}_i(\boldsymbol{\alpha})$  no corresponda a la verdadera matriz de correlación de  $\mathbf{y}_i$  (Davis 2002, pág. 300).

Como  $\widehat{V}_{EEG}$  es un estimador consistente de la varianza de  $\widehat{\alpha}_{EEG}$  entonces se tiene que  $\widehat{V}_{EEG} \xrightarrow{P} \mathbb{V}(\widehat{\alpha}_{EEG})$ , es decir, para cualquier  $\epsilon > 0$  y para cualquier  $\eta > 0$ , existe un entero positivo  $n_0 \equiv n_0(\epsilon, \eta)$ , tal que

$$P \left[ \left\| \widehat{V}_{EEG} - \mathbb{V}(\widehat{\alpha}_{EEG}) \right\| > \epsilon \right] < \eta \text{ siempre que } n \geq n_0.$$

Sea  $\epsilon^*$  tal que

$$\epsilon = \frac{\epsilon^*}{\|\Phi\| \|\Phi^T\|}$$

por lo que

$$P \left[ \|\widehat{V}_{EEG} - \mathbb{V}(\widehat{\alpha}_{EEG})\| > \epsilon \right] < \eta \text{ siempre que } n \geq n_0,$$

es equivalente a

$$P \left[ \|\Phi \widehat{V}_{EEG} \Phi^T - \mathbb{V}(\Phi \widehat{\alpha}_{EEG})\| > \epsilon^* \right] < \eta \text{ siempre que } n \geq n_0,$$

y por tanto  $\widehat{\mathbb{V}}(\widehat{\beta}_{EEG}(t)) \xrightarrow{P} \mathbb{V}(\widehat{\beta}_{EEG}(t))$ .