
Modelos de coeficientes dinámicos generalizados para el análisis de datos longitudinales en salud: una aplicación en HIV/AIDS y COVID-19

Generalized time-varying coefficient models for the analysis of longitudinal data in health: an application in HIV / AIDS and COVID-19

Elkin Javier Casadiego Rincón.^a
elkincasadiego@usantotomas.edu.co

Juan Camilo Sosa Martínez^b
jcsosam@unal.edu.co

Resumen

El análisis de datos longitudinales es necesario cuando la variable respuesta se mide repetidamente sobre la misma unidad de observación a lo largo del tiempo. Los métodos *paramétricos* se han empleado tradicionalmente en el análisis de datos longitudinales para estimar los coeficientes que definen la relación entre el predictor lineal y la variable respuesta, sin embargo las técnicas paramétricas no son apropiadas cuando no se cumplen los supuestos acerca de la variable respuesta y la componente aleatoria del modelo, o cuando el valor esperado de la variable respuesta (o una función de esta variable vía una función de enlace) no resulta ser una función conocida de los efectos fijos y aleatorios, razones por las que los modelos paramétricos pueden llevar a conclusiones alejadas de la tendencia promedio del conjunto de datos. En estos casos, las técnicas de regresión *no paramétricas*, en las que en lugar de parámetros se emplean funciones locales suavizadas que dependen del tiempo, denominados coeficientes o parámetros dinámicos, constituyen una alternativa muy poderosa de modelamiento en el análisis de datos longitudinales, puesto que permiten establecer una dependencia funcional más flexible entre la variable respuesta y las covariables. Este trabajo propone desarrollar técnicas de estimación e inferencia para modelos de coeficientes dinámicos no paramétricos generalizados, particularmente cuando la variable respuesta es de conteo, ilustrando su aplicación en el efecto de la carga viral sobre el conteo de células CD4, en pacientes con HIV/AIDS sometidos a un tratamiento antirretroviral, y también en la predicción de casos de COVID-19.

Palabras clave: Análisis de datos longitudinales, funciones de base radial kernel, regresión spline, modelos de coeficientes dinámicos, carga viral, conteo de linfocitos CD4, HIV/AIDS, COVID-19.

^aEstudiante

^bDirector

Abstract

Longitudinal data analysis is necessary when the response variable is repeatedly measured on the same observation unit over time. The *parametric* methods have been traditionally used in the analysis of longitudinal data to estimate the coefficients that define the relationship between the linear predictor and the response variable. However, parametric techniques do not work when the assumptions about the response variable and the random component of the model are not fulfilled, or when the expected value of the response variable (or a function of this variable via a link function) is not be a known function of the fixed and random effects, reasons why parametric models can draw conclusions away from the average trend of the data set. In these cases, *non-parametric* regression techniques, in which time-dependent smoothed local functions are used instead of parameters, called coefficients or dynamic parameters, constitute a very powerful modeling alternative in the analysis of longitudinal data, since they allow establish a more flexible functional dependence between the response variable and the covariates. In this work, it is proposed to develop estimation and inference techniques for generalized non-parametric dynamic coefficient models, particularly when the response variable is counting, illustrating its application in the effect of viral load on CD4 cell count, in patients with HIV / AIDS undergoing antiretroviral treatment, and also in the prediction of COVID-19 cases.

Keywords: Longitudinal data analysis, radial basis kernel function, regression spline, time-varying coefficient model, viral load, CD4 T lymphocytes count, HIV/AIDS, COVID-19

Índice

1. Introducción	5
2. Modelos de coeficientes dinámicos	8
2.1. Modelo lineal de coeficientes dinámicos	8
2.2. Modelo lineal generalizado de coeficientes dinámicos	9
2.2.1. Estimación mediante regresión Spline	9
2.2.2. Estimación utilizando Kernel como funciones base	11
2.2.3. Estimación de los parámetros de suavizamiento	12
2.2.4. Inferencia estadística bootstrap	13
3. Contexto clínico y epidemiológico de los casos a estudiar	15
3.1. Contexto HIV/AIDS	15
3.2. Contexto pandemia COVID-19	16
4. Aplicaciones	19
4.1. Caso de estudio ACTG-388	19
4.2. Caso de estudio ACTG-315	23
4.3. Aplicación caso COVID-19.	30
5. Discusión y Conclusiones	32
6. Notación	33

Índice de figuras

1.	Gráfica en "spaghetti" de la población y tendencia lineal de algunos individuos del conteo de linfocitos CD4 a través del tiempo con los datos ACTG-388.	20
2.	Tendencia lineal del conteo de linfocitos CD4 en el tiempo con los datos ACTG-388. . . .	20
3.	Intercepto (β_0) del conteo de linfocitos CD4 en ACTG-388.	22
4.	Bandas de confianza del intercepto del conteo de linfocitos CD4 en el tiempo de seguimiento con los datos ACTG-388 en escala natural, se comparan las metodologías spline y kernel. . . .	22
5.	Gráfica en "spaghetti" de la población y tendencia lineal de algunos individuos del conteo de linfocitos CD4 a través del tiempo con los datos ACTG-315.	24
6.	Tendencia lineal del conteo de linfocitos CD4 a través del tiempo en ACTG-315.	24
7.	Gráfica en "spaghetti" de la población y tendencia lineal de algunos individuos del efecto de la carga viral sobre el conteo de linfocitos CD4 con los datos ACTG-315.	25
8.	Tendencia lineal del conteo de linfocitos CD4 frente a la carga viral en ACTG-315.	25
9.	Tendencia del conteo de linfocitos CD4 frente a la carga viral con ACTG-315, escogiendo en algunos puntos en el tiempo de seguimiento.	26
10.	Intercepto (β_0) en valores originales y pendiente (β_1) del conteo de linfocitos CD4 en ACTG-315 en el tiempo con kernel.	27
11.	Intercepto (β_0) dado en valores originales y pendiente (β_1) del conteo de linfocitos CD4 en ACTG-315 en el tiempo con splines.	27
12.	Modelo Kernel del conteo de linfocitos CD4 con los datos ACTG-315.	28
13.	Modelo Spline del conteo de linfocitos CD4 con los datos ACTG-315.	28
14.	Modelo Spline para los datos de COVID-19 en Colombia.	30
15.	Modelo kernel para los datos de COVID-19 en colombia.	31

1. Introducción

El análisis de datos longitudinales (LDA, por sus siglas en inglés) es de interés cuando se observa repetidamente un conjunto de n individuos a través del tiempo, registrándose los valores de una variable respuesta junto con las respectivas covariables, que pueden depender o no del instante en que se han medido. En tal caso, las observaciones de un mismo sujeto no son independiente, considerándose las mediciones entre los individuos independientes (Twisk, 2013). Se puede modelar la variable respuesta estimándose la tendencia promedio poblacional o la evolución de cada individuo, caso en el que estaríamos hablando de un modelo mixto o aleatorio.

La relación entre la variable respuesta y las covariables en un LDA se puede modelar a través de un enfoque *paramétrico*, especificando completamente la distribución de la variable respuesta por medio de un conjunto constante y finito de coeficientes, que extienden naturalmente el modelo de regresión lineal de datos transversales (Davidian, 2007). En los modelos paramétricos los coeficientes tienen una interpretación inmediata y las técnicas de estimación usuales como mínimos cuadrados ordinarios/generalizados, máxima verosimilitud y máxima verosimilitud restringida aplican directamente (Cuevas, 2004). Dos de estas técnicas son las ecuaciones de estimación generalizadas (GEE, por sus siglas en inglés) y el modelo lineal de efectos mixtos (LME, por sus siglas en inglés). En las GEE se asume que la variable respuesta puede tener o no distribución normal y se tiene en cuenta directamente la posible correlación de las mediciones dentro de cada sujeto, a través de la denominada matriz de correlación de trabajo. A diferencia de las GEE, el LME modela explícitamente la variación dentro de los individuos ajustando una distribución normal alrededor de unos parámetros de regresión determinados y estimando la varianza de esta distribución, lo cual permite modelar directamente las características propias de cada individuo (Twisk, 2013, ver capítulo 4, para mas detalles).

Las desventajas potenciales de los modelos *paramétricos* ocurren cuando no se cumplen los supuestos acerca de la variable respuesta y la componente aleatoria del modelo, o cuando la variable respuesta (o una función de esta variable vía una función de enlace) no resulta ser una función conocida de los efectos fijos y los efectos aleatorios, razones por las que los modelos paramétricos pueden llevar a conclusiones alejadas de la tendencia promedio del conjunto de datos. En estos casos, las técnicas de regresión *no paramétricas* constituyen una alternativa muy poderosa de modelamiento. Un tipo de modelo no paramétrico que permite realizar LDA son los modelos de coeficientes dinámicos (TVCM, por sus siglas en inglés de time-varying coefficient models), en los cuales en lugar de parámetros se emplean funciones locales suavizadas que dependen del tiempo, denominados coeficientes o parámetros dinámicos, y permiten establecer una dependencia funcional más flexible entre la variable respuesta y las covariables (Sosa and Díaz, 2012; Lu and Huang, 2017; Wu and Tian, 2018, pág 14). La estrategia no paramétrica no presupone la especificación del modelo, y es especialmente útil cuando se quiere captar la dependencia lineal de la variable respuesta a unas covariables que dependen del tiempo, sin embargo presenta debilidades cuando las covariables presentan alta dimensionalidad (Fan and Zhang, 2008).

Una forma usual de modelar la respuesta media en un modelo TVCM es asumir que los errores aleatorios siguen un proceso Gaussiano. Sin embargo algunos autores como Huang et al. (2012); Lu and Huang (2017), han señalado una desventaja en trabajos precedentes (Sosa and Díaz, 2009, 2012; Huang and Lu, 2016), en los que han asumido una distribución simétrica (como la normal) para la variable respuesta ante eventuales comportamientos sesgados de los errores. Por lo tanto, la inferencia estadística basada sobre este supuesto de normalidad puede llevar a resultados imprecisos.

En algunas ocasiones los problemas de asimetría en los errores se pueden arreglar realizando transformaciones en la variable respuesta, por ejemplo linealizando aplicando escala log. Cuando la función de respuesta parezca ser curvilínea, se pueden realizar transformaciones para mantener al modelo como de primer orden. Un modelo de orden menor en una variable transformada puede ser preferible a un modelo de orden superior en la métrica original (Montgomery et al., 2006). Pero si un polinomio de bajo orden proporciona mal ajuste a los datos, y aumentar el orden del polinomio no mejora mucho la situación, se deben emplear soluciones alternativas a la transformación de la variable respuesta. Este problema se puede presentar cuando la función se comporta en forma distinta en diferentes partes del rango de

la covariable. Una forma bien conocida de solución cuando la variable respuesta es sesgada consiste en aplicar la transformación de Box-Cox, permitiendo que la variable transformada tenga una distribución condicional Gaussiana, como lo realizaron en un LDA, Lipsitz et al. (2000). Una limitación de esta aproximación es que la forma de la transformación es invariante en el tiempo. Para un LDA, la variable respuesta $Y(t)$ puede tener diferentes sesgos sobre el tiempo t , de tal manera que una transformación fija puede no llevar a una aproximación Gaussiana para todo t dentro del rango de interés (Wu and Tian, 2018, pag. 443). Una manera de solucionar esto es teniendo en cuenta la naturaleza de la variable con un modelo generalizado y realizando ajustes no paramétricos en las covariables como las que se proponen en este trabajo.

En consecuencia, una extensión de TVCM para una variable de respuesta no-normal, es un modelo generalizado introduciendo una función de enlace conocida que relacione el predictor dinámico lineal y la variable respuesta. En este contexto, entre otros desarrollos Hastie and Tibshirani (1993) propusieron realizar *varying-coefficient models* en que las covariables pueden depender o no del tiempo, los coeficientes son funciones suavizadas y la respuesta media esta unida al predictor lineal vía una función de enlace, el modelo se ejemplificó para una respuesta binaria; Wu and Zhang (2006) desarrollaron modelos de coeficientes dinámicos generalizados con una función de enlace logística para datos longitudinales de repuesta binaria, por sus características lo llamó *generalized nonparametric population mean (GNPM) model*; en Fan and Zhang (2008) se propusieron métodos estadísticos para *generalized varying-coefficient models* incluyendo aquellos para datos longitudinales; Lu and Zhang (2009) desarrollaron un *generalised varying-coefficient mixed model* usando coeficientes dinámicos para estimar la función media; Şentürk and Müller (2008) desarrollaron un modelo de coeficientes dinámicos generalizados para datos longitudinales teniendo en cuenta rezagos en el predictor lineal; Park and Jeong (2018) desarrollaron un *Poisson autoregressive varying-coefficient model*, introduciendo una estructura de autocorrelación e ilustrándolo con datos diarios de homicidios; ver también a Wang (2007) para otros ejemplos.

A pesar de los desarrollos citados en modelos de coeficiente dinámicos no paramétricos generalizados, los trabajos previos publicados por Sosa and Díaz (2009, 2012); Sosa and Buitrago (2021), no tuvieron en cuenta posibles asimetrías en el comportamiento de los errores en las aplicaciones analizadas. Por lo anterior, se propone en este trabajo soluciones para superar las limitaciones planteadas. Por una parte, se asume que la variable respuesta es un conteo y se aplica una distribución de Poisson para modelar el valor esperado de dicha variable; por otro lado, se emplea un modelo lineal generalizado para estimar los coeficientes dinámicos, y se usa una función de enlace conocida sobre la variable respuesta.

Los modelos que se desarrollan en el presente trabajo se ejemplifican estableciendo la relación dinámica de dos variables en el contexto del VIH/SIDA y en la predicción del número de casos que puedan ocurrir en la pandemia COVID-19. El primer ejemplo muestra la importancia de usar los TVCM en el curso de la respuesta a los tratamientos para monitorizar pacientes en el contexto clínico, y el segundo permite apreciar la aplicabilidad de la metodología en el contexto epidemiológico, en donde interesa realizar predicciones de los casos, importante por sus implicaciones en la toma de decisiones en salud pública.

Los datos que se emplean en las ilustraciones para el caso VIH/SIDA son de dos estudios clínicos sobre el SIDA desarrollados por el AIDS Clinical Trials Group¹. En el primero denominado ACTG-315 se incluyeron a 46 pacientes infectados con HIV-1 y tratados con terapia antirretroviral consistente de ritonavir, zidovudine, lamivudine. Tanto la carga viral como el conteo de CD4 fueron medidos simultáneamente durante los días 0, 2, 7, 10, 14, 28, 56, 84, 168, y 196 (Lederman et al., 1998). En el segundo estudio denominado ACTG-388 (Fischl et al., 2003) incluyeron a 517 pacientes con enfermedad avanzada, y fueron tratados con indinavir mas efavirenz o nelfinavir, en las sección sobre las aplicaciones se darán mas detalles sobre estas datas. Los datos para el ejemplo COVID-19 en Colombia se obtienen de los repositorios oficiales de las páginas gubernamentales².

¹Visitar <https://actgnetwork.org/> para mas información acerca del grupo.

²<https://www.datos.gov.co/api/views/gt2j-8ykr/rows.csv?accessType=DOWNLOAD>

El objetivo del presente trabajo es desarrollar técnicas de estimación de coeficientes dinámicos para datos de conteo, a través de regresión Spline y funciones radiales Kernel (RKF, por sus siglas en inglés) en el contexto de un modelo lineal generalizado, que permita caracterizar la evolución del efecto de la carga viral sobre el conteo de células CD4 en pacientes con VIH/SIDA sometidos a un tratamiento antirretroviral, y también se ejemplifica la metodología pronosticando el número de casos que ocurren en un periodo de tiempo específico en la actual pandemia de la COVID-19 en Colombia. Así mismo, se desarrollan métodos de inferencia sobre los coeficientes de los modelos vía técnicas de remuestreo, y finalmente se evalúa el desempeño de la estimación en términos de precisión y eficiencia a través de medidas de bondad de ajuste y de error.

El documento se organiza inicialmente en la sección 2.1 presentando los componentes teóricos de los modelos lineales de coeficientes dinámicos, en la sección 2.2 se desarrolla el modelo no paramétrico generalizado que se empleará en las aplicaciones y las estrategias de suavizamiento mediante Spline y kernel. Luego en la sección 3 se documenta el contexto de los ejemplos en salud que emplearemos para ilustrar los modelos que se aplicarán. Y finalmente en las dos últimas secciones se presentan los resultados, la discusión y conclusiones.

2. Modelos de coeficientes dinámicos

2.1. Modelo lineal de coeficientes dinámicos

Un modelo semiparamétrico para LDA en una aplicación para el conteo de células CD4 fue propuesto por Zeger and Diggle (1994). Una generalización del modelo propuesto por Zeger and Diggle (1994) donde los coeficientes varían con el tiempo (TVCM) fue el desarrollado por Hoover et al. (1998). El cual es de la forma:

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}(t_{ij}) + \epsilon_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, n, \quad (2.1)$$

donde n es el número de individuos en estudio, n_i es la cantidad de mediciones del i -ésimo individuo, t_{ij} es el instante del tiempo en que se realizó la j -ésima medición del i -ésimo individuo, y_{ij} es el valor de la variable respuesta asociada con el i -ésimo individuo tomada en el tiempo t_{ij} , $\mathbf{x}_{ij} \equiv \mathbf{x}_{ij}(t_{ij})$, $\mathbf{x}_i(t_{ij}) = [x_{i0}(t_{ij}), \dots, x_{id}(t_{ij})]^T$ es el vector de $(d+1) \times 1$ covariables medidas en el instante t_{ij} correspondientes al i -ésimo individuo, cuyas componentes pueden variar o no en el tiempo, $\boldsymbol{\beta}(t_{ij}) = [\beta_0(t_{ij}), \beta_1(t_{ij}), \dots, \beta_d(t_{ij})]^T$ es el vector de $(d+1) \times 1$ coeficientes dinámicos, también denominados parámetros dinámicos, evaluados en el tiempo t_{ij} . Por su parte, $\epsilon_{ij} \equiv \epsilon_{ij}(t_{ij})$ denota el error relacionado con la medición del i -ésimo individuo en el instante t_{ij} , y se asume como un proceso estocástico con media cero, independiente de $\mathbf{x}_{ij}$ y con función de covarianza $\delta(t, t') = Cov[\epsilon(t), \epsilon(t')]$, además se supone que las mediciones entre individuos son independientes (Şentürk and Müller, 2008). En un principio no hay restricciones fuertes sobre las $(d+1)$ funciones que conforman el vector de coeficientes dinámicos, solo que sean lo suficientemente suaves (derivables), de tal modo que sea posible la aplicación de los procedimientos de estimación (Sosa and Díaz, 2009).

Cuando en el modelo 2.1 no se considera alguna covariable, es decir con $d = 0$, y $x_{i0}(t_{ij}) = 1$, para todo $i = 1, \dots, n$, el modelo puede escribirse como:

$$y_{ij} = \beta_0(t_{ij}) + \epsilon_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, n; \quad (2.2)$$

el modelo 2.2 se empleará con la aplicación ACTG-388, teniendo en cuenta que esta data no cuenta con covariables, y también para la aplicación COVID-19.

Si en el modelo 2.1 solo se considera una covariable dependiente del tiempo, es decir con $d = 1$, y $x_{i0}(t_{ij}) = 1$, para todo $i = 1, \dots, n$, el modelo puede escribirse como:

$$y_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij})x_1(t_{ij}) + \epsilon_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, n; \quad (2.3)$$

el cual se aplicará a la data ACTG-315 que contiene la covariable carga viral.

En un TVCM la variable respuesta puede ser cuantitativa continua o discreta, y en este último caso se podría modelar por medio de un proceso Gaussiano cuando la distribución discreta es simétrica. El modelo expuesto hasta ahora está diseñado para trabajar con variables cuantitativas que se asumen con distribución normal, lo cual no es siempre pertinente por las razones que se expusieron en la introducción, y por tal motivo es necesario hacer una extensión del modelo que tenga en cuenta la naturaleza de otro tipo de variables, como las variables de conteo.

2.2. Modelo lineal generalizado de coeficientes dinámicos

Una extensión de un TVCM para una variable de respuesta no-normal, es un modelo generalizado introduciendo una función de enlace conocida que relacione el predictor dinámico lineal y la variable respuesta.

En un modelo lineal generalizado para LDA la media poblacional marginal y la varianza de la respuestas y_{ij} son especificadas como:

$$E(y_{ij} | t_{ij}) = \mu_{ij}, \quad Var(y_{ij} | t_{ij}) = \phi w_{ij}^{-1} V(\mu_{ij}) \quad j = 1, 2, \dots, n_i, \quad i = 1, 2, \dots, n,$$

donde ϕ es un parámetro de escala, w_{ij} son los pesos, y $V(\cdot)$ es una función conocida de la varianza, representa la dependencia de la varianza de la respuesta a su media. En un modelo de coeficientes dinámicos generalizados la media marginal μ_{ij} es asumida estar relacionada con el predictor lineal $\mathbf{x}_{ij}^T \boldsymbol{\beta}(t_{ij})$ a través de una función de enlace $g(\cdot)$. El modelo estaría expresado (Cai et al., 2000; Wang, 2007; Wu and Zhang, 2006, Cap. 10) de la siguiente forma:

$$g(\mu_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta}(t_{ij}) \quad j = 1, 2, \dots, n_i; \quad i = 1, 2, \dots, n. \quad (2.4)$$

En nuestro experimento se considera el peso de cada individuo $w_i = 1/(nn_i)$, con $i = 1, \dots, n$, donde cada sujeto es inversamente ponderado por su número de medidas n_i , de tal manera que los individuos con pocas medias repetidas recibe mas peso que los sujetos con mas medidas repetidas.

A continuación se describen dos métodos de estimación no paramétricos de las funciones $\beta_r(t)$, $r = 0, 1, \dots, d$ a partir de funciones bases (Wu and Zhang, 2006; Ramsay et al., 2009). Una estimación global mediante regresión spline (RS) y otra estimación local con funciones radiales kernel (RKF, por su sigla en inglés de Radial kernel Function) (Wu and Tian, 2018).

2.2.1. Estimación mediante regresión Spline

Una *regresión Spline* es una función que expresa cada coeficiente dinámico $\boldsymbol{\beta}(t)$ en el modelo 2.4 como una combinación lineal de funciones base de Spline (basis spline) (Wu and Tian, 2018, Pag. 97). Una base de spline (Φ) es cada uno de los términos o bloques de que se compone la combinación lineal de funciones base (también denominada expansión de funciones base) (Ramsay et al., 2009, Pag. 29). En otras palabras, la regresión Spline se considera como un polinomio a trozos de orden $g + 1$, ó de grado g , $g \geq 0$. El término grado se refiere a la potencia más alta en el polinomio (Wu and Zhang, 2006, Pag. 50).

De manera específica, un polinomio a trozos es una función de valor real $f(\cdot)$ definida en el rango de interés, esto es sobre un intervalo $[a, b]$, el cual esta dividido en un número de particiones, $a = \tau_0 < \tau_1 < \tau_2 < \dots < \tau_K < \tau_{K+1} = b$, formando subintervalos contiguos conocidos como nodos ($\tau_r, r = 1, 2, \dots, K$), en donde el número de nodos es dado por K (Wu and Zhang, 2006, Pag. 50).

Para representar las funciones base de Splines se emplean distintas técnicas, entre otras se encuentran: la base de polinomios, M-splines, I-splines, la base de potencias truncadas (TPB, por sus siglas en inglés de truncated power basis), y base de B-splines (BSB, por sus siglas en inglés de B-splines basis), (Wu and Zhang, 2006; Ramsay et al., 2009, Pag.44).

Un TPB es un conjunto de funciones bases de splines, el cual puede representarse (Wu and Zhang, 2006, Pag. 51) usando g grados y K nodos así:

$$1, t, \dots, t^g, (t - \tau_1)_+^g, \dots, (t - \tau_K)_+^g, \quad (2.5)$$

donde $(t - \tau_K)$ denota w , y $w_+^g = (w_+)^g$ es la potencia g de la parte positiva de w , con $w_+ = \max(0, w)$.

En resumen, al conjunto 2.5 de $K + g + 1$ funciones base se le conoce como base de potencias truncadas de orden $g + 1$ ó de grado g , con nodos interiores $\tau_\ell, \ell = 1, 2, \dots, K$.

Usando el conjunto 2.5, es posible expresar una regresión spline, es decir cada coeficiente dinámico $\beta_r(t), r = 0, 1, \dots, d$, como:

$$\beta_r(t) = \sum_{s=0}^g \alpha_{r,s} t^s + \sum_{\ell=1}^K \alpha_{r,g+\ell} (t - \tau_K)_+^g, \quad (2.6)$$

donde $\alpha_r = (\alpha_{r,0}, \alpha_{r,1}, \dots, \alpha_{r,K_r+g+1})$ son $p_r \times 1$ vector columna de coeficientes desconocidos asociados, siendo $p_r = K_r + g + 1$ el número de funciones base.

Ahora, si por conveniencia se representase las TPB (2.5) como:

$$\Phi_{r,p_r}(t) = [1, t, \dots, t^g, (t - \tau_1)_+^g, \dots, (t - \tau_K)_+^g]^T,$$

denotando $\Phi_{r,p_r}(t)$ las funciones base involucradas evaluadas en el tiempo t , y p_r un número entero no negativo indicando el número de funciones base, entonces la regresión spline $\beta_r(t)$ (2.6) puede ser reexpresada (Wu and Zhang, 2006, Pag. 279) así:

$$\beta_r(t) = \Phi_{r,p_r}(t)^T \alpha_r \quad (2.7)$$

Si $\beta_r(t)$ tal como está expresado en 2.7, se sustituye en 2.4, el modelo de coeficientes dinámicos generalizado incluyendo las covariables, quedaría:

$$g(\mu_{ij}) = \mathbf{z}_{ij}^T \alpha \quad j = 1, 2, \dots, n_i, \quad i = 1, 2, \dots, n, \quad (2.8)$$

donde $y_{ij} \equiv y_i(t_{i,j})$, y $\mathbf{z}_{i,j} = (\mathbf{z}_{0,i,j}^T, \dots, \mathbf{z}_{d,i,j}^T)$, con

$$\mathbf{z}_{r,i,j} = x_{r,i}(t_{i,j}) \Phi_r(t_{i,j}) \quad r = 0, 1, \dots, d.$$

Similar a α cada $\mathbf{z}_{ij}$ es un vector columna de la covariable veces el vector de funciones base de $p \times 1$. Para el i -ésimo sujeto $i = 1, 2, \dots, n$, se denota el vector respuesta y la matriz de diseño como:

$$\mathbf{y}_i = (y_{i,1}, \dots, y_{i,n_i}), \quad \mathbf{Z}_i = [\mathbf{z}_{i,1}, \dots, \mathbf{z}_{i,n_i}]^T,$$

en consecuencia, el vector respuesta y la matriz de diseño para todos los datos se denota así:

$$\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_n^T), \quad \mathbf{Z} = [\mathbf{Z}_1^T, \dots, \mathbf{Z}_n^T],$$

lo cual permite expresar a 2.8 en forma matricial como un modelo lineal generalizado estándar:

$$g(\boldsymbol{\mu}) = \mathbf{Z}\alpha, \quad (2.9)$$

donde $\boldsymbol{\mu}$ es el valor esperado del vector de respuestas $\mathbf{y}$, esto es, $\boldsymbol{\mu} = E(\mathbf{y})$. Una vez estimado α usando un apropiado método, es directo conseguir una estimación de $\beta(t)$.

2.2.2. Estimación utilizando Kernel como funciones base

Los Kernels son funciones que se asocian a cada uno de los datos, entonces la suma ponderada de estas funciones es un estimador para aproximar la función de densidad desconocida. La definición del estimador está determinado por 2 parámetros: el ancho de banda (h), y la función núcleo (Wu and Tian, 2018). La función núcleo K , define la forma de los pesos que se asocian a cada observación dentro de la banda definida, y por tanto, determina su importancia en el cálculo de la estimación. Para dicha función es habitual el uso de alguna de las siguientes RKF:

$$\begin{aligned} \text{Triangular} : k(x) &= (1 - |x|), & |x| \leq 1 \\ \text{Epanechnikov} : k(x) &= 3/4(1 - x^2), & |x| \leq 1 \\ \text{Biponderado} : k(x) &= 15/16(1 - x^2)^2, & |x| \leq 1 \\ \text{Gaussiano} : k(x) &= \left(\frac{1}{\sqrt{2\pi}} \right) \exp(-x^2/2), & x \in (-\infty, +\infty) \end{aligned}$$

El ancho de banda h , se define como un parámetro positivo cuyo rango en principio será la amplitud del intervalo de estimación. No obstante, la elección de dicho parámetro constituye uno de los aspectos cruciales del procedimiento de estimación. Las buenas propiedades del estimador resultante dependerán en gran medida de la elección que se haga de dicho parámetro. En este sentido existe una compensación (trade-off en inglés) entre flexibilidad y complejidad que es controlada a través de dicho parámetro. Así, si dicho parámetro se toma muy pequeño, tan solo observaciones muy próximas al punto de estimación intervendrán en el cálculo del estimador, describiendo muy bien comportamientos locales, pero obtendremos una curva estimada muy variable. Si, por el contrario, dicho parámetro se toma muy grande, las estimaciones en cada punto se verán afectadas por observaciones en puntos alejados, de forma que difícilmente se podrán recoger los comportamientos locales, dando lugar a grandes sesgos y como consecuencia obtendremos poca variabilidad. Desde un punto de vista teórico, la elección del ancho de banda tendrá que buscar una adecuada compensación entre sesgo y varianza.

La estrategia de estimación a través de RKF, es aproximar cada componente de $\beta(t)$ mediante RKF tratándolas como funciones radiales base (RBF, por su sigla en inglés). Esto lleva a modelar $\beta_r(t)$, $r = 0, 1, \dots, d$, mediante la expresión:

$$\beta_r(t) = \sum_{\ell=1}^M \xi_r \left(\frac{|t - t_\ell|}{h} \right) \alpha_{r\ell} = \Xi_r(t)^T \alpha_r, \quad (2.10)$$

donde M es la cantidad de tiempos de medición diferentes t_ℓ , $|\cdot|$ denota el valor absoluto, y ξ_r es alguna función kernel, h es un ancho de banda y $\alpha_r = [\alpha_{r1}, \dots, \alpha_{rM}]$ es el vector $M \times 1$ de coeficientes asociados. $\Xi_r(t)$ es un vector de $M \times 1$ conformado por las funciones kernel utilizadas en la aproximación,

$$\Xi_r(t) = \left[\xi_r \left(\frac{|t - t_1|}{h} \right), \dots, \xi_r \left(\frac{|t - t_M|}{h} \right) \right]^T \quad (2.11)$$

Si $\beta_r(t)$ tal como está expresado en 2.10, se sustituye en 2.4, el modelo de coeficientes dinámicos generalizado incluyendo las covariables, quedaría:

$$g(\mu_{ij}) = \mathbf{v}_{ij}^T \alpha \quad j = 1, 2, \dots, n_i, \quad i = 1, 2, \dots, n, \quad (2.12)$$

donde $y_{ij} \equiv y_i(t_{i,j})$, y $\mathbf{v}_{i,j} = (\mathbf{v}_{0,i,j}^T, \dots, \mathbf{v}_{d,i,j}^T)$, con

$$\mathbf{v}_{r,i,j} = x_{r,i}(t_{i,j}) \Xi_r(t_{i,j}) \quad r = 0, 1, \dots, d.$$

Similar a α cada $\mathbf{v}_{ij}$ es un vector columna de la covariable veces el vector de funciones base de $p \times 1$. Para el i -ésimo sujeto $i = 1, 2, \dots, n$, se denota el vector respuesta y la matriz de diseño como:

$$\mathbf{y}_i = (y_{i,1}, \dots, y_{i,n_i}), \quad \mathbf{V}_i = [\mathbf{v}_{i,1}, \dots, \mathbf{v}_{i,n_i}]^T,$$

en consecuencia, el vector respuesta y la matriz de diseño para todos los datos se denota así:

$$\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_n^T), \quad \mathbf{V} = [\mathbf{V}_1^T, \dots, \mathbf{V}_n^T],$$

lo cual permite expresar a 2.12 en forma matricial como un modelo lineal generalizado estándar:

$$g(\boldsymbol{\mu}) = \mathbf{V}\boldsymbol{\alpha}, \quad (2.13)$$

donde $\boldsymbol{\mu}$ es el valor esperado del vector de respuestas $\mathbf{y}$, esto es, $\boldsymbol{\mu} = E(\mathbf{y})$. Al igual que con spline una vez estimado α usando un apropiado método, es directo conseguir una estimación de $\beta(t)$.

El procedimiento usado para estimar los coeficientes α , ya sea en 2.9 ó en 2.13, una vez establecidos los nodos y el ancho de banda óptimos, es empleando máxima verosimilitud por el método de mínimos cuadrados iterativamente ponderados (IRLS, por la sigla en inglés de iteratively reweighted least squares) que está por defecto en la función *glm.fit* del paquete *glm* en R.

2.2.3. Estimación de los parámetros de suavizamiento

La selección del número de nodos y el ancho de banda (los parámetros de suavizamiento ($\boldsymbol{\rho}$)) es un paso crítico en el proceso de estimación. El vector de parámetros necesario usando regresión spline es la selección de los nodos $\boldsymbol{\rho} = [p_0, \dots, p_d]^T$, mientras que el vector de parámetros empleando RKF, además de la selección de los nodos, se requiere la selección del ancho de banda $\boldsymbol{\rho} = [h_0, \dots, h_d]^T$. La selección de estos parámetros debe ser adecuadamente realizado porque el ajuste del modelo depende de la mejor elección.

Los parámetros de suavizamiento están relacionados directamente con el conjunto de funciones base con que se expresan los coeficientes dinámicos. En el caso de trabajar con TPB, los parámetros de suavizamiento están dados por

$$\boldsymbol{\rho} = K_r + g_r + 1, \quad r = 0, 1, \dots, d, \quad (2.14)$$

donde K_r es el número de nodos y g_r es el grado de la base (polinomio) asociada con la estimación del r -ésimo coeficiente dinámico. Seleccionar el valor de los parámetros de suavizamiento es seleccionar los K_r y los g_r . Seleccionar el grado del polinomio consiste en elegir si es de 1, 2 o 3 grado, es decir si es lineal, cuadrático o cúbico, siendo que la mayoría de referencias prefieren el cúbico. Sin embargo para elegir el número de nodos si se requiere de un criterio establecido por métodos adecuados.

Ubicación de los nodos

En regresión mediante spline una vez fijado el número de nodos K_r a través de algún criterio apropiado, se deben ubicar en el rango de interés, que en el caso de los datos longitudinales es a lo largo del conjunto de los tiempos igualmente espaciados en el intervalo temporal de interés $[a; b]$. Los valores que definen este intervalo son el mínimo y el máximo de todos los tiempos, así

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

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

En la práctica para espaciar los nodos, los tiempos de medición se escalan de tal manera que el valor mínimo sea 0 y el máximo 1. Lo que permite segmentar el dominio de la covariable tiempo y operar los parámetros de suavizamiento (numero de nodos y ancho de banda).

Selección de la cantidad de nodos y del ancho de banda

En el caso de modelos normales, se suele seleccionar el número de nodos y el ancho de banda óptimos por medio de la validación cruzada excluyendo un punto (PCV) (sigla en ingles de leave-one-point-out cross-validation), propuesto en Wu and Zhang (Wu and Zhang, 2006, pág 286), y previamente implementado por Hoover D. R. (Hoover et al., 1998, pág 814). El método PCV permite seleccionar el mejor parámetro de suavizado sin tener en cuenta la estructura de correlación en las mediciones dentro de cada individuo (Wu and Zhang, 2006, pág 287). La idea es elegir el vector de parámetros de suavizamiento que minimiza la expresión:

$$PCV(\rho) = \sum_{i=1}^n \sum_{j=1}^{n_i} w_i \left[y_{ij} - \mathbf{x}(t_{ij})^T \hat{\boldsymbol{\beta}}^{(-ij)}(t_{ij}) \right]^2, \quad (2.15)$$

donde $w_i, i = 1, \dots, n$, es el peso asociado con el i -ésimo individuo, $w_i = 1/(nn_i)$ es el esquema de pesos elegido y $\hat{\boldsymbol{\beta}}^{(-ij)}(t_{ij})$ es una estimación de $\boldsymbol{\beta}(t_{ij})$ usando todos los datos excepto la j -ésima medida del i -ésimo individuo.

Considerando que $\mathbf{A}$ es la matriz suavizada por las respuesta en todos los puntos de diseño, definida como $\mathbf{A} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$, donde $\mathbf{X}$ es la matriz de funciones base spline o kernel, tal que $\hat{\mathbf{y}} = \mathbf{A}\mathbf{y}$, se ha demostrado que 2.15 es equivalente a la siguiente expresión según (Wu and Zhang, 2006, pág 286),

$$PCV(\rho) = \frac{(\mathbf{y} - \mathbf{A}\mathbf{y})^T \mathbf{W}(\mathbf{y} - \mathbf{A}\mathbf{y})}{(1 - \text{tr}(\mathbf{A})/N)^2}, \quad (2.16)$$

donde $\mathbf{y}$ es el vector de la variable respuesta, $N = \sum_{i=1}^n n_i$ es el número total de observaciones en todos los individuos, $\mathbf{W} = \text{diag}(w_1 \mathbf{I}_{n_1}, \dots, w_n \mathbf{I}_{n_n})$, $\mathbf{W}_i = w_i \mathbf{I}_{n_i}$ es la matriz de pesos del i -ésimo individuo o unidad experimental, $i = 1, \dots, n$, $\mathbf{I}_k$ es la matriz identidad de $k \times k$, y $\text{tr}(\mathbf{A})$ denota la traza de $\mathbf{A}$.

En nuestra experiencia, aunque el anterior método suele ser robusto a desviaciones de la normalidad, otro método que empleamos para seleccionar el número de nodos y el ancho de banda óptimos fue el criterio de información de Akaike (AIC), que establece una relación entre la bondad de ajuste y su complejidad. El AIC esta dado por la formula $2p - 2\ln(L)$, donde p es el número de parámetros evaluados del modelo y L es el máximo valor de la función de verosimilitud para el modelo estimado. El mejor modelo es el que presenta menor AIC. Para estos propósitos se construyó un algoritmo que permitía seleccionar los parámetros óptimos tanto para $\boldsymbol{\alpha}_0$ como para $\boldsymbol{\alpha}_1$ de manera distinta.

2.2.4. Inferencia estadística bootstrap

Los intervalos de confianza de una medición puntual de un parámetro pueden ser establecidos mediante la técnica de remuestreo *bootstrap*. Para el caso no paramétrico como se tiene es una función de estimación de los coeficientes se desarrolla una banda de confianza. Asumiendo independencia en los individuos, se realiza un remuestreo de la totalidad de las mediciones repetidas de cada sujeto con reemplazamiento a partir de los datos originales (Wu and Tian, 2018, Pag.73). Para obtener la función de estimación por remuestreo *bootstrap* se pueden seguir los siguientes pasos:

- Selección aleatoria de n sujetos con reemplazamiento a partir de la muestra original, del mismo tamaño de la muestra.
- La muestra longitudinal *bootstrap* obtenida por reemplazamiento es denotada por:

$$[(\mathbf{Y}_{ij}^*, t_{ij}^*) : i = 1, \dots, n; j = 1, \dots, n_i],$$

algunos valores de la muestra original pueden aparecer de manera repetida en la la nueva muestra *bootstrap*.

- c. Obtener la matriz de diseño aplicando el suavizador no paramétricos sobre la muestra *bootstrap* obtenida.
- d. Estimar los $\alpha_r = (\alpha_{r,0}, \alpha_{r,1}, \dots, \alpha_{r,K_r+g+1})$ para spline ó los $\alpha_r = [\alpha_{r,1}, \dots, \alpha_{r,M}]$ para kernel, con la información obtenida en los pasos anteriores, aplicando la función *glm.fit* del paquete *glm* en R.
- e. Calcular la curva de estimadores no parametricos basado en la muestra *bootstrap* denominado estimador $\hat{\mu}_k^{boot}(t; h; w)$ de $\mu(t)$.
- f. Repetir los pasos anteriores B veces. Se denota el B bootstrap estimator $b = 1, \dots, B$, a $\hat{\mu}_k^{boot}(t; h; w)$ de $\mu(t)$ se tal manera que los estimadores obtenidos de la B muestras son:

$$B_\mu^B(t; h, w) = \hat{\mu}_k^{boot,1}(t; h; w), \dots, \hat{\mu}_k^{boot,B}(t; h; w)$$

- g. Las muestras *bootstrap* generadas en la secuencia anterior se usan para aproximar los intervalos o bandas de confianza para $\mu(t)$, basados en los percentiles de las muestras *bootstrap*, lo cual está dado por:

$$(\mathbf{L}_{(\alpha/2)}(t), \mathbf{U}_{(\alpha/2)}(t)),$$

donde $\mathbf{L}_{(\alpha/2)}(t)$ y $\mathbf{U}_{(\alpha/2)}(t)$ son los $(\alpha/2)$ y $(1 - \alpha/2)$, bajo y superior percentil de $B_\mu^B(t; h, w)$, respectivamente.

3. Contexto clínico y epidemiológico de los casos a estudiar

3.1. Contexto HIV/AIDS

El virus de la inmunodeficiencia humana tipo 1 (HIV-1, por sus siglas en inglés) es un retrovirus (de la familia Retroviridae), que contiene información genética bajo la forma de ácido ribonucleico (RNA, por sus siglas en inglés), y causa una infección que para la mayoría de los individuos no tratados disminuye los Linfocitos T (LT) CD4, resultando en morbilidad y mortalidad (Gonzalo-Gil et al., 2017). El deterioro progresivo del sistema inmunológico por la infección del HIV-1, produce en los estadios avanzados el síndrome de inmunodeficiencia adquirida (AIDS, por su sigla en inglés), el cual se define por la presencia de alguna de las más de 20 infecciones oportunistas o de cánceres relacionados con el HIV (WHO, 2007). El RNA cuantitativo del HIV (carga viral) y el conteo de LT CD4, son dos de los marcadores pronósticos de progresión de los pacientes con HIV/AIDS comúnmente usados, y juegan un importante rol en la valoración inicial y en la evaluación de la respuesta al tratamiento (Ministerio, 2014).

La carga viral es útil para determinar el pronóstico, establecer la línea de base para evaluar la respuesta al tratamiento y vigilar la supresión viral (Hughes et al., 1997). Su medición regular se ha relacionado con resultados clínicos importantes, orienta la evaluación de resistencia al tratamiento, infecciones oportunistas, mortalidad, y es un marcador de progresión de la enfermedad (Marschner et al., 1998; Thiébaud et al., 2000). Se ha establecido que el nivel de replicación viral, medido por el número de copias de RNA por *mL* (mililitro) o por su logaritmo (en base 10), determina el tiempo desde la infección inicial a AIDS y muerte, dado que un elevado valor inicial es un marcador de rápida progresión, y en todos los individuos la progresión clínica es precedido por un incremento en el RNA (Korenromp et al., 2009).

Debido a que con el progreso de la infección se observa una disminución en el conteo de LT CD4, su determinación es un parámetro inmunológico que ayuda a establecer el riesgo de complicaciones asociadas al HIV, la necesidad de profilaxis para infecciones oportunistas y la indicación de tratamiento con antirretrovirales (Ministerio, 2014; Gonzalo-Gil et al., 2017). Al comienzo de una terapia antirretroviral activa, el conteo de LT CD4 demostró ser el factor predictivo más importante de progresión clínica; así, en un estudio los pacientes que iniciaron el tratamiento con menos de 200 células/ μL (microlitro) tuvieron sustancialmente mayor riesgo de progresión clínica que aquellos con conteos mayores (Egger et al., 2002). El conocimiento de la relación de la carga viral y del conteo de LT CD4, su valor predictivo y su participación en la historia natural de la infección, han incrementado en importancia debido a que desde una perspectiva poblacional, permiten establecer los umbrales para el inicio de los tratamientos y las profilaxis de manera más eficiente, y desde una perspectiva individual, permiten monitorizar la respuesta al tratamiento y prever progresión de la enfermedad (Korenromp et al., 2009). El valor pronóstico de los marcadores, varía dependiendo de algunos factores como son: el estado de la enfermedad, la edad, condición del paciente (embarazo, comorbilidades, estado inmunológico de base), el efecto directo del virus independiente del impacto sobre los LT CD4, el desconocimiento del tiempo de la infección a la seroconversión, y de la duración del seguimiento de los estudios, entre los más importantes (Korenromp et al., 2009).

Es bien conocido que los LT CD4 son blancos del HIV y declinan con la progresión de la infección, de tal manera que cuando la terapia antiviral suprime la carga viral, el conteo de LT CD4 puede recuperarse hasta niveles normales (Lederman et al., 1998). A pesar del amplio consenso sobre el carácter predictivo de la carga viral y del conteo de LT CD4, se ha observado que su relación es cambiante, tanto en ausencia como en presencia de tratamiento antirretroviral, lo que amerita desarrollar modelos apropiados que permitan caracterizar esta relación en el tiempo (Liang et al., 2003). La variabilidad del conteo de LT CD4 y de la carga viral dentro y entre los individuos en los tiempos medidos, el carácter multivariado que los relaciona, así como las distintas metodologías empleadas para su análisis, dificultan su modelado conjunto para establecer riesgo o pronóstico (Korenromp et al., 2009). En general, se ha observado que la respuesta virológica (medida a través de la carga viral) y la respuesta inmunológica (medida a través del conteo de los LT CD4), están negativamente correlacionadas durante el tratamiento antiviral (Ford et al., 2015). Sin embargo, como esta relación es influenciada por diversos factores, puede no ser

constante durante todo el tratamiento y no todos los pacientes logran la supresión viral esperada, lo que es importante conocer para una adecuada monitorización de las respuestas a los tratamientos. Algunos estudios han confirmado que la relación entre la carga viral y el conteo de LT CD4 tienen diferentes patrones, lo que implica que pueden existir discordancias entre las respuestas virológicas e inmunológicas que deben ser evaluadas tanto en la población como individualmente (Liang et al., 2003; Korenromp et al., 2009).

Por lo expuesto, un análisis longitudinal de los marcadores es más apropiado que un análisis transversal, permitiendo monitorear la dinámica de la relación con mayor información, dado que esta estrategia permite tener en cuenta factores dinámicos relevantes para el análisis. Por ejemplo, uno de los supuestos de la regresión lineal para datos transversales consiste en que las mediciones de la variable respuesta constituyen un conjunto de observaciones independientes, por el contrario, las observaciones efectuadas sobre un mismo sujeto a través del tiempo suelen ser dependientes unas de otras, lo que amerita ser tenido en cuenta en el desarrollo de un modelo longitudinal.

Para modelar la relación poblacional entre la carga viral y el conteo de LT CD4, Sosa y Díaz (2009) propusieron utilizar un TVCM donde los errores aleatorios siguen un proceso Gaussiano. Sin embargo, como lo han señalado Huang et al., 2012 y Lu and Huang, 2017, una desventaja de la propuesta de Sosa y Díaz (2009) consiste en que el modelo asume una distribución simétrica, como la normal, para la variable respuesta, lo cual no es apropiado para eventuales comportamientos sesgados de los errores. Por lo tanto, la inferencia estadística basada sobre este supuesto de normalidad puede llevar a resultados engañosos. Otro elemento importante de distinción con trabajos precedentes, es que no se empleará la carga viral como variable respuesta (Sosa and Díaz, 2009, 2012; Huang and Lu, 2016; Lu and Huang, 2017), sino al conteo de LT CD4, teniendo en cuenta que por la historia natural de la enfermedad se sabe que los LT CD4 son uno de los principales blancos del HIV. Esta decisión plantea la necesidad de proponer un modelo dinámico no paramétrico generalizado con respuesta de conteo.

3.2. Contexto pandemia COVID-19

La enfermedad por Coronavirus 2019 (COVID-19, por su sigla en inglés de Coronavirus Disease 2019), es el nombre dado a la patología causada por la infección del SARS-CoV-2 (por su sigla en inglés de Severe Respiratory Acute Syndrome Coronavirus 2), un virus RNA que hace parte de la subfamilia de virus Orthocoronavirinae, género Betacoronavirus (Millán-Oñate et al., 2020; Trujillo, 2020). Los coronavirus (CoV) son virus que surgen periódicamente en diferentes áreas del mundo y que causan Infección Respiratoria Aguda (IRA). La infección por el SARS-CoV-2 ha sido catalogado por la Organización Mundial de la Salud (WHO, por sus siglas en inglés) como una emergencia en salud pública de importancia internacional (ESPII).

La enfermedad fue detectada en diciembre de 2019, en pacientes que habían estado expuestos posiblemente a transmisión alimentaria desde animales salvajes en un mercado de la ciudad de Wuhan, provincia de Hubei, China, siendo reconocida semanas después en enero del 2020 (Li et al., 2020). La WHO inició a reportar casos de COVID-19 el 21 de enero del 2020, luego de lo cual la epidemia se diseminó detectándose casos en todos los continentes, y el 6 de marzo se confirmó el primer caso en Colombia. La infección se produce cuando una persona es inoculada con el virus eliminado a partir de una persona infectada sintomática o no que expulsa partículas del virus (Trujillo, 2020).

Estimar el curso de la epidemia es una de las tareas mas importantes en el control de la enfermedad, dado que permite evaluar la magnitud de afectación en la población y el impacto de las medidas de mitigación. Desde que la COVID-19 se generalizó en la población mundial, esfuerzos por establecer la dinámica de la pandemia han surgido con iniciativas académicas y gubernamentales. Para poder predecir epidemias se emplean herramientas matemáticas que permiten crear modelos epidemiológicos determinísticos o estocásticos. Los modelos determinísticos a diferencia de los estocásticos consideran a los individuos pertenecientes a un estado del modelo (Manrique Abril et al., 2020). Estos modelos están centrados en los estados por los que pueden pasar los individuos, tratando a los individuos de manera

conjunta. Al estudiar modelos centrados en estados, cobra especial importancia la variación de cada estado, representado por la derivada de los mismos (de Pereda Sebastián et al., 2010, pag.16).

Diversos modelos para establecer el comportamiento de la COVID-19 se han propuesto desde el inicio de la epidemia, comenzando por los clásicos modelos compartimentados como el SIR, del acrónimo de susceptible-infectados-recuperados, que usa tres ecuaciones diferenciales ordinarias para describir la dinámica del flujo de las personas entre los tres compartimentos (Luo, 2020, pag. 5). Los actuales modelos compartimentados desarrollados para COVID-19 son versiones del trabajo original de Kermack and McKendrick (1927). Estos modelos son paramétricos, en los que se debe definir las condiciones iniciales a partir de parámetros extraídos de experiencias previas, por lo que su precisión esta determinada por los ajustes periódicos o calibraciones que se haga a las condiciones de modelado.

Se asume que las epidemias en su fase de crecimiento tienen un comportamiento exponencial (Ma, 2020; Kasilingam et al., 2020), que se modifica por la historia natural de la enfermedad o por las medidas de mitigación o prevención instauradas para su control (confinamiento, aislamientos, vacunas, etc). Estimar la tasa de crecimiento de una curva epidémica, es fundamental para caracterizar su dinámica y es un desafío porque decae con el tiempo (Kasilingam et al., 2020). Es por esta razón que estos procesos son generalmente analizados por medio de *curvas de crecimiento* que tienen generalmente forma *sigmoidal*(ς), y describen cómo el proceso incrementa en un intervalo de tiempo hasta un nivel límite o nivel de saturación (Kumar, 2015). A diferencia de la función exponencial que no tiene tope, con las curvas en ς se debe especificar su máximo.

Hay una amplia variedad de curvas de crecimiento con forma sigmoidal aplicados para modelar la COVID-19, de las cuales, las que utilizan la ecuación logística, uno de los modelos clásicos en dinámica de poblaciones, son muy empleados. Algunos ejemplos que emplean curvas de crecimiento son: el trabajo de Kasilingam et al. (2020) explora el crecimiento de los casos de COVID-19 en 42 países usando modelos exponenciales y de aprendizaje automático, incluida la regresión logística, para predecir signos de contención temprana; el trabajo de Ma (2020) compara un modelo SIR con uno exponencial, logístico y con el modelo de Richards (una modificación del modelo logístico); Pelinovsky et al. (2020) aplicó una ecuación logística generalizada para interpretar el comportamiento de la epidemia y encontró que los datos en promedio eran bien descritos; Wang et al. (2020) integraron los datos epidemiológicos de COVID-19 antes del 16 de junio de 2020 en el modelo logístico para ajustar el límite de la tendencia epidémica y luego introdujeron este valor límite en un modelo predictivo de series de tiempo basado en aprendizaje automático para derivar la curva de la epidemia y predecir su tendencia; Batista (2020) estimaron el tamaño final de la epidemia por un modelo logístico, y alternativamente desarrollaron un modelo basado en la función Weibull para el mismo objetivo; ejemplo distinto a la aplicación de la función logística fue el que Ahmadi et al. (2020) presentaron al emplear otros modelos sigmoidales como lo son el Gompertz y el von Bertalanffy para predecir el COVID-19 en Iran.

A diferencia de los modelos paramétricos, los modelos empíricos incluyendo las regresiones suavizadas, no requieren especificar las condiciones iniciales porque dependen exclusivamente de los datos. En esta línea encontramos los trabajos entre otros, de Ekum and Ogunsanya (2020) que aplicaron una regresión polinomial para predecir los casos de COVID-19; Barbosa et al. (2020) emplearon una regresión *Joinpoint* para estimar la incidencia acumulada de la pandemia en algunos países y realizar comparaciones entre las curvas; Díaz-Narváez et al. (2020) exploraron el mejor tipo de curva o modelo de tendencia para explicar el comportamiento epidemiológico del contagio por COVID-19, los datos fueron recogidos del Departamento de epidemiología del Ministerio de Salud de Chile y emplearon cuatro modelos diferentes: cuadrático, exponencial, exponencial suavizado simple y exponencial suavizado doble, el que mejor se ajusto a los datos fue una curva doble exponencial suave; Clara-Rahola (2020) realizaron una aproximación empírica a la dinámica de los casos diagnosticados y definieron escenarios, encontrando dos comportamientos exponenciales; Carroll et al. (2020) emplearon un enfoque de análisis funcional de datos tipo FPCA dinámicos (por su siglas en ingles de Functional Principal Component Analysis) para hacer predicciones a corto plazo de trayectorias futuras; finalmente mencionaremos el trabajo de Mercker et al. (2020) quienes en el contexto de regresiones suavizadas usó GAMM (por sus siglas en ingles de generalised additive mixed model) con una estructura de autocorrelación de orden uno, porque el numero de infectados

están temporalmente autocorrelacionados, para explicar y predecir el comportamiento de la epidemia en Alemania y sus distritos. En esta línea de modelos empíricos con regresiones suavizadas, emplearemos el modelo de coeficientes dinámicos generalizados que desarrollaremos, asumiendo el caso colombiano como un caso especial para predecir el comportamiento inicial de la epidemia y en un corto tiempo.

4. Aplicaciones

4.1. Caso de estudio ACTG-388

El *AIDS Clinical Trials Group* (ACTG, por sus siglas en inglés), fundado en 1987, es una de las redes de ensayos clínicos para HIV más grandes del mundo. Este primer ejemplo es basado en un estudio del ACTG, el protocolo 388³, que fue un ensayo aleatorizado, abierto en 517 pacientes con HIV/AIDS avanzado (elegibles si el conteo de células CD4 ≤ 200 células/ μ L o un nivel plasmático de RNA HIV-1 ≥ 80.000 copias/ mL), distribuidos en tres grupos a recibir tratamiento antiretroviral altamente activo con lamivudinaa mas zidovudina e indinavir (grupo indinavir) que fué el grupo base de comparación, frente a efavirenz (grupo efavirenz mas indinavir) o nelfinavir (grupo nelfinavir mas indinavir), y fueron monitorizados por 2,1 años (Fischl et al., 2003).

En esta aplicación utilizamos los datos de uno de los grupos de tratamiento del protocolo ACTG-388, que incluye 166 pacientes tratados por 120 semanas, durante las cuales el conteo de células CD4 fue monitorizado. Esta base de datos no cuenta con otras covariables y presentamos el comportamiento del conteo de células CD4 en el tiempo. Algunos pacientes no se adhieren al esquema de seguimiento y ocurre pérdida de mediciones de la variable respuesta en algunas semanas, lo que es frecuente en este tipo de estudios longitudinales, resultando en una data no balanceada. El número de mediciones por paciente tiene un rango de 1 a 18 observaciones, en total son 2107 observaciones, y el conteo de CD4 tiene un rango de 0 a 1364 células/ μ L.

Inicialmente presentamos una descripción del conjunto de datos para definir la trayectoria global del conteo de CD4 en el tiempo, usando metodologías convencionales, mostrando lo limitado que estos métodos resultan para datos longitudinales.

En la figura 1 al lado izquierdo, se visualiza una gráfica tipo “spaghetti”, que representa las trayectorias de todos los individuos en la variable respuesta frente al tiempo en semanas, sin definirse una clara tendencia poblacional. Un análisis escogiendo algunos individuos, presentado al lado derecho de la misma figura, evidencia que algunos presentan una tendencia creciente, otros decreciente ó sin tendencia. Sin embargo, estas visualizaciones no resultan esclarecedoras para entender cuál es el comportamiento promedio del conteo de linfocitos CD4 a través del tiempo, por lo que se realiza un ajuste lineal al conjunto de puntos, permitiendo observar una tendencia global creciente del conteo de linfocitos CD4 con el tiempo, figura 2. A pesar de ello esta metodología no es suficiente para mostrar en detalle la dinámica o tendencia del conteo de linfocitos CD4 en cada punto del tiempo, información necesaria para monitorizar continuamente la respuesta de los pacientes al tratamiento antirretroviral, y para tomar decisiones clínicas.

El modelo de coeficientes que varían con el tiempo no paramétrico y generalizado para esta aplicación es de la forma:

$$\log(E(CD4_{ij})) = \beta_0(t_{ij}) \quad j = 1, \dots, n_i, \quad i = 1, \dots, n, \quad (4.17)$$

donde $CD4_{ij}$ es el conteo de linfocitos CD4 que se asume una variable con distribución Poisson, asociado con la j -ésima medida del i -ésimo paciente, y $\beta_0(t_{ij})$ es el parámetro dinámico asociado con el intercepto expresado como en 2.10 o como en 2.7, según se estime mediante kernel o spline, respectivamente.

En el desarrollo del TVCM generalizado propuesto para la data 388, se emplea el criterio PCV para la selección del número óptimo de parámetros de suavizamiento, tal como está previamente descrito en la sección 2.2.3. Con este criterio para la estimación del intercepto y siguiendo la metodología spline, se seleccionaron 6 nodos, y con la metodología kernel usando también 6 nodos se obtuvo que el mejor modelo es el que emplea un ancho de banda de 0.29.

³<http://www.urmc.rochester.edu/biostat/people/faculty/wusite/datasets/data/ACTG388Data1Arm.cfm>.

Una vez establecidos los parámetros óptimos, se estimaron los coeficientes α empleando máxima verosimilitud por el método de mínimos cuadrados iterativamente ponderados (IRLS, por la sigla en inglés de iteratively reweighted least squares), que está por defecto en la función *glm.fit* del paquete *glm* en R.

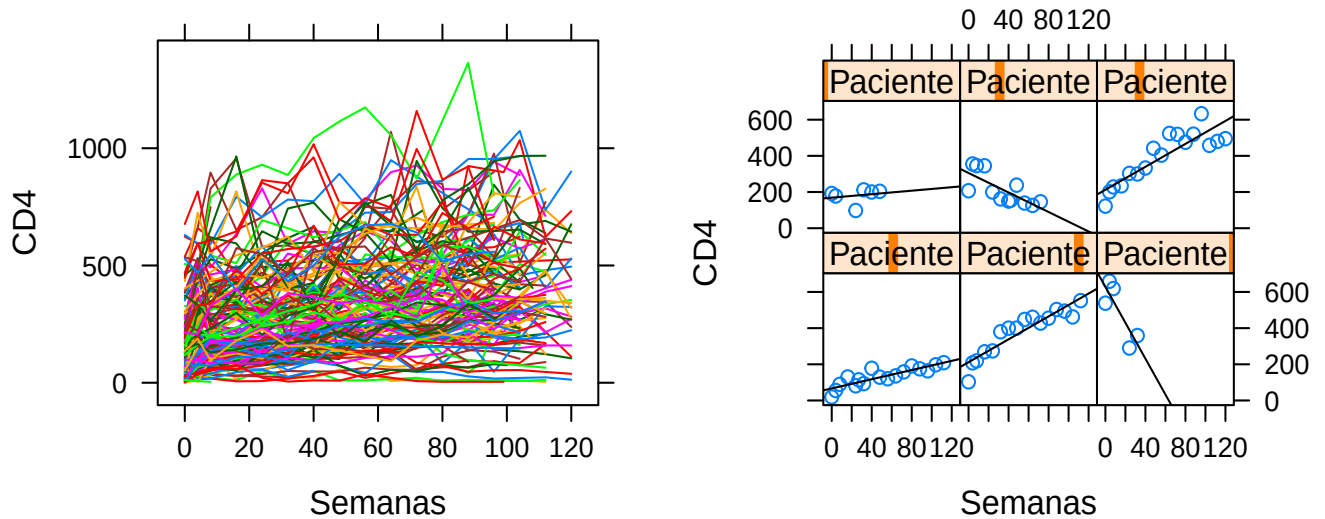


Figura 1: Gráfica en “spaghetti” de la población y tendencia lineal de algunos individuos del conteo de linfocitos CD4 a través del tiempo con los datos ACTG-388.

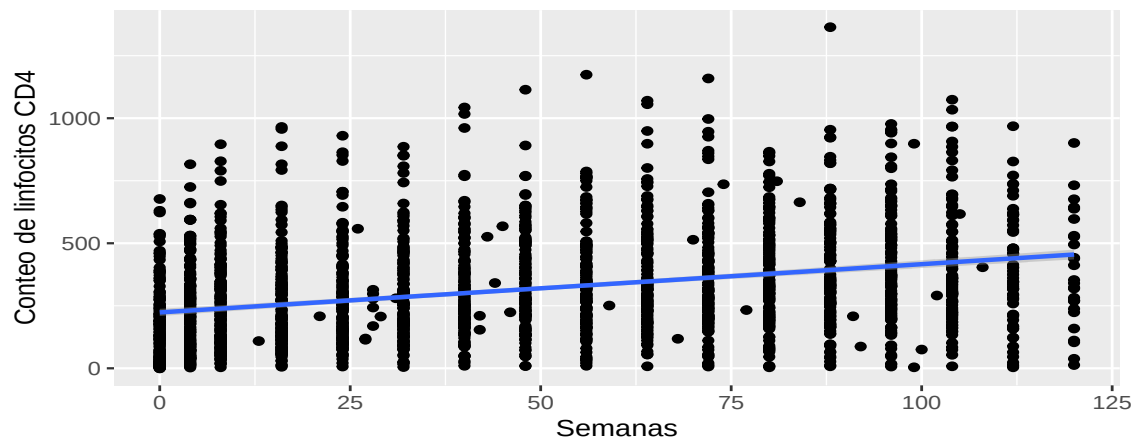


Figura 2: Tendencia lineal del conteo de linfocitos CD4 en el tiempo con los datos ACTG-388.

Para evaluar la calidad de los modelos propuestos se emplean diversas estrategias. Por una parte, se usa la deviance residual, que permite evaluar indirectamente la adecuación del modelo en la estimación de los alfas, verificando además si hay problemas de dispersión. La AIC, que establece una relación entre la bondad de ajuste y la complejidad del modelo, permite escoger el de mejor desempeño entre las estrategias splines y kernel estudiadas. También se usa la raíz cuadrada del error cuadrático medio (RMSE, por su sigla en inglés) que es una evaluación directa sobre la bondad de ajuste de las respuestas estimadas, la cual permite establecer que tan distantes están los valores predichos de los observados, usándose para este fin la función *rmse* del paquete *metrics* en R. Los modelos con menores valores del AIC y del RMSE son los escogidos.

Sobre la *deviance residual* (residual deviance) se efectúa una prueba de bondad de ajuste, empleando la distribución χ^2 como criterio de decisión, según la cual al compararse el tamaño de la deviance residual frente a sus grados de libertad se espera sean iguales en un modelo sin problemas de dispersión. Lo anterior permite establecer la hipótesis nula que la deviance residual sigue una distribución χ^2 , y un valor no significativo en la prueba indica que no hay evidencia de una pérdida de ajuste del modelo o problemas de dispersión (Faraway, 2016, Pag.158). Los resultados permiten evidenciar que no existen problemas de sobredispersión en los métodos, conclusión que se extrae de la prueba χ^2 sobre la deviance residual, la cual permite tomar la decisión de no rechazar la hipótesis nula, según la cual el modelo es adecuado. La bondad de ajuste sobre la deviance residual se presenta como $p(\chi^2)$ en la tabla 1.

Con el fin de extraer los coeficientes α , se probaron diversas distribuciones para modelar la variable respuesta en el modelo lineal generalizado. La binomial negativa (bn) presenta el mejor desempeño dado por el menor AIC, tanto con Kernel como para Spline. Así mismo, se aplicó la prueba de Vong para seleccionar el mejor modelo, bajo la hipótesis nula según la cual los modelos son indistinguibles (Cameron and Trivedi, 2013, Pag.198). Los resultados con el test de Vong corroboran la interpretación inicial en donde una binomial negativa es mejor opción que usar una Poisson, sin embargo no se distingue en su desempeño entre la metodología kernel y la Spline usando en ambos una binomial negativa. En la tabla 1 se presentan los resultados de evaluación de la calidad y ajuste de los modelos.

Modelo	AIC	RMSE	$p(\chi^2)$
Kernel Poisson	151.313	199.7203	1
Kernel bn	27.17015	199.9836	1
Spline Poisson	156.1622	199.4498	1
Spline bn	33.16135	199.5138	1

Tabla 1: Desempeño del modelo generalizado para estimar α con los datos ACTG-388.

El $\beta_0(t)$ obtenido para la data ACTG-388 con ambos métodos se presenta a continuación en la figura 3.

Para construir las bandas de confianza alrededor de la función de estimación de los coeficientes dinámicos, se realizó un remuestreo de la totalidad de las mediciones repetidas de cada sujeto con reemplazamiento a partir de los datos originales, y se obtuvieron n muestras (1000) con las que se efectuaron las estimaciones de los α , el valor central de los β y luego sus límites superior e inferior (percentiles 2.5 % y 97.5 %, respectivamente), tal como se describió en la sección sobre inferencia bootstrap. En la figura 4 se presentan las bandas de confianza del intercepto al 95 % de credibilidad tanto por los métodos Spline y Kernel.

Ambos modelos exhiben comportamientos similares, excepto porque con Spline se observa que la trayectoria media del conteo de CD4 presenta un crecimiento pronunciado o abrupto inicialmente, con un máximo hacia la semana 15 (289 células/ μ L), para luego descender continuamente hasta la semana 26 (262 células/ μ L), desde lo cual la tendencia continúa creciendo pero con fluctuaciones hasta la semana 101 (404 células/ μ L), y a partir de ahí desciende nuevamente pero lentamente hasta el final del periodo. El comportamiento con la metodología kernel es más suave pero creciente hasta la semana 112 (404 células/ μ L), desde la cual se presenta un decrecimiento hasta el final. En general el crecimiento es más pronunciado en las primeras 40 semanas de tratamiento y continúa incrementándose pero más despacio entre las semanas 101 ó 112, según el método empleado, a partir del cual el conteo declina.

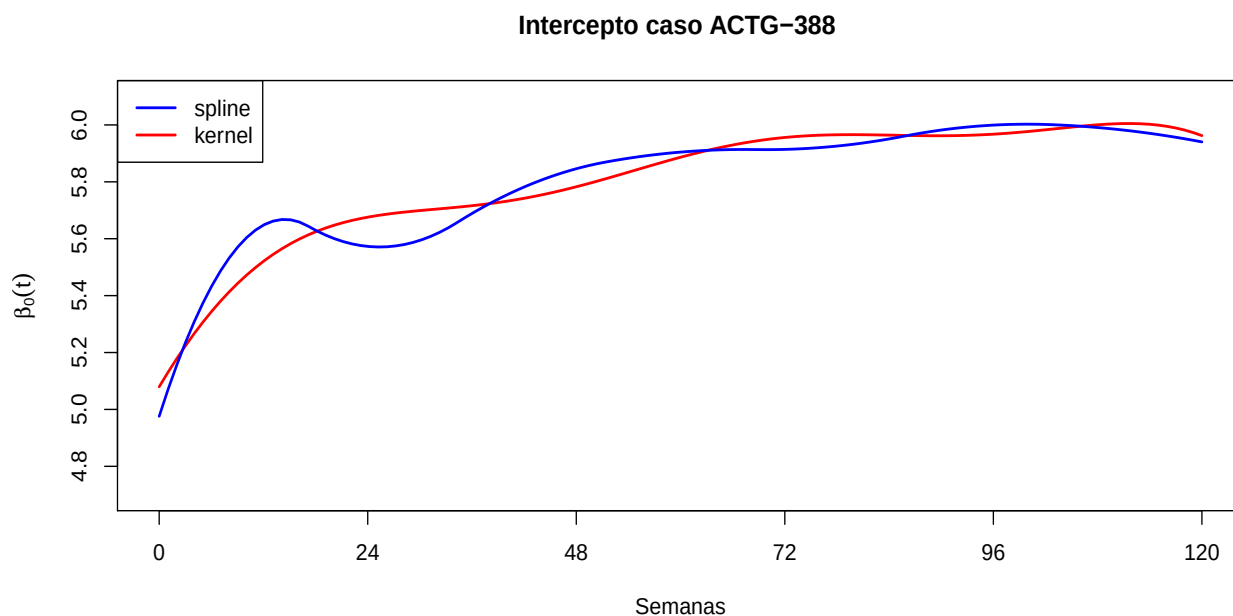


Figura 3: Intercepto (β_0) del conteo de linfocitos CD4 en ACTG-388.

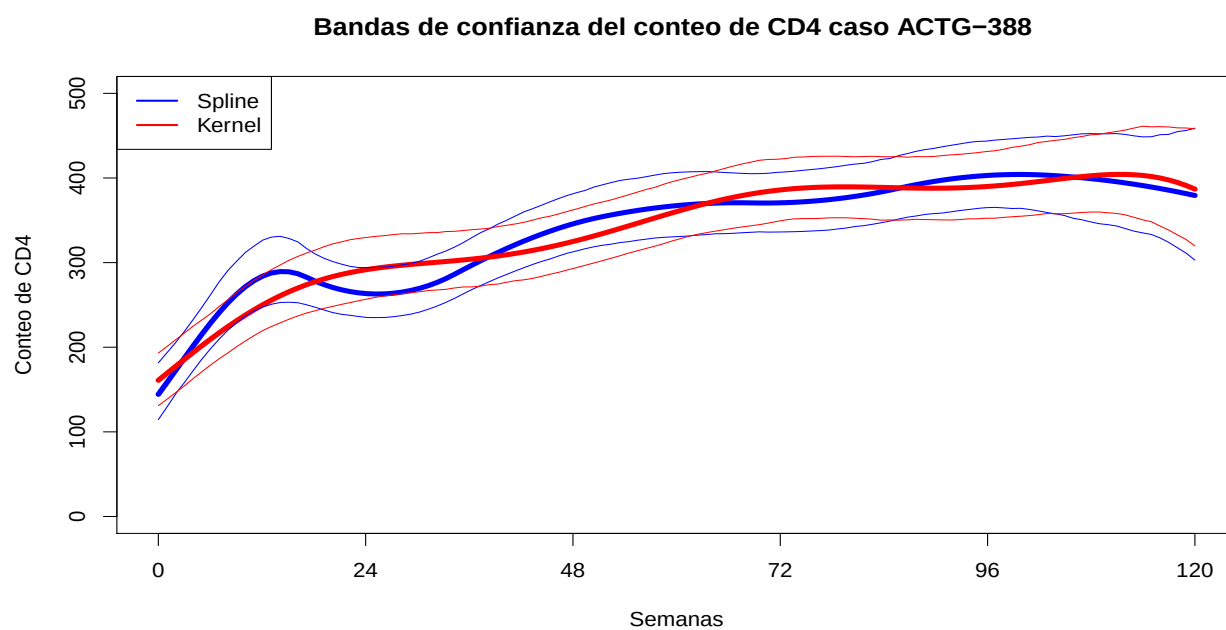


Figura 4: Bandas de confianza del intercepto del conteo de linfocitos CD4 en el tiempo de seguimiento con los datos ACTG-388 en escala natural, se comparan las metodologías spline y kernel.

En resumen, los anteriores resultados podrían señalar un efecto del tratamiento hasta casi el final del tiempo de seguimiento, con decline en las últimas semanas.

Resultados similares fueron encontrados por Wu and Zhang (2006) empleando spline con un polinomio de grado 2 y 4 nodos. Observaron que la función media estimada de la población aumentó drásticamente durante las primeras semanas, y continuó aumentando a un ritmo más lento hasta aproximadamente la semana 110, para luego reducirse hasta el final del estudio. En el análisis de residuales para evaluar el adecuado ajuste del modelo a los datos, encontró que los residuales estandarizados no eran simétricos con un importante sesgo, por lo que recomendó alguna transformación en los datos (Wu and Zhang, 2006, Pag. 124).

En un trabajo reciente Sosa and Buitrago (2021) modelando la data ACTG-388 encontraron resultados similares dado porque el conteo medio de las células CD4 se incrementó rápidamente durante las primeras 40 semanas de tratamiento, y continúa creciendo lentamente hasta la semana 100, luego de lo cual cae hasta el fin del tratamiento. Los análisis de los residuales indicó que el modelo ajusta adecuadamente. Aplicando el Deviance Information Criterio (DIC) como medida de bondad de ajuste concluyen que el TVCM basada en funciones kernel es el modelo que mejor desempeño presenta.

4.2. Caso de estudio ACTG-315

El protocolo ACTG-315 ⁴ evalúa el efecto de la terapia antirretroviral intensiva sobre la respuesta inmune. Consiste de 12 semanas administrando la combinación de Zidovudine, Lamivudine, y Ritonavir. El tratamiento se probó en 46 individuos adultos mayores de 18 años, con moderada infección por VIH-1 (es decir con un conteo de linfocitos CD4 entre 100 y 300 células/ μ L), de los cuales 44 recibieron tratamiento al menos 9 semanas de las 12 que duró el estudio (Lederman et al., 1998, p. 72).

En específico, la data analizada aquí presenta un tiempo máximo de seguimiento de 196 días y cada paciente cuenta con un número distinto de medidas repetidas, las cuales variaron de 4 a 10, registrándose en total 361 observaciones. Teniendo en cuenta que no todos los individuos presentan el mismo número de mediciones los datos no están balanceados. La carga viral esta dada en escala logarítmica, y el conteo de linfocitos CD4 esta en el rango de 17 a 488 células/ μ L.

Al igual que con el caso anterior, realizamos inicialmente una descripción de los datos empleando metodologías convencionales para luego desarrollar los análisis con los modelos de coeficientes dinámicos generalizados propuestos.

En el lado izquierdo de la figura 5 se presenta el comportamiento de todos los individuos del conteo de linfocitos CD4 a través del tiempo en una representación conocida como en "spaghetti". Al lado derecho se muestran algunos individuos, encontrando que la tendencia es creciente en algunos y en otros es decreciente.

En la figura 6 se muestra que la tendencia promedio en la población del conteo de linfocitos CD4 a través del tiempo es creciente, sin embargo esta aproximación no permite visualizar aun la dinámica en cada punto.

En la figura 7 se presenta una visión poblacional del comportamiento del conteo de linfocitos CD4 en relación con la carga viral, pareciendo indicar una relación inversa. Sin embargo, como se muestra al lado derecho de la gráfica, la tendencia individual es fluctuante y algunos tienen una tendencia positiva y otros una tendencia nula.

En la figura 8 se muestra una regresión lineal sobre todos los puntos, que exhibe efectivamente una relación inversa entre las variables. Detalle este que es importante monitorizar en cada punto por sus implicaciones en el seguimiento y control clínico de los pacientes.

⁴<http://www.urmc.rochester.edu/biostat/people/faculty/wusite/datasets/data/ACTG315LongitudinalDataViralLoadData.cfm>.

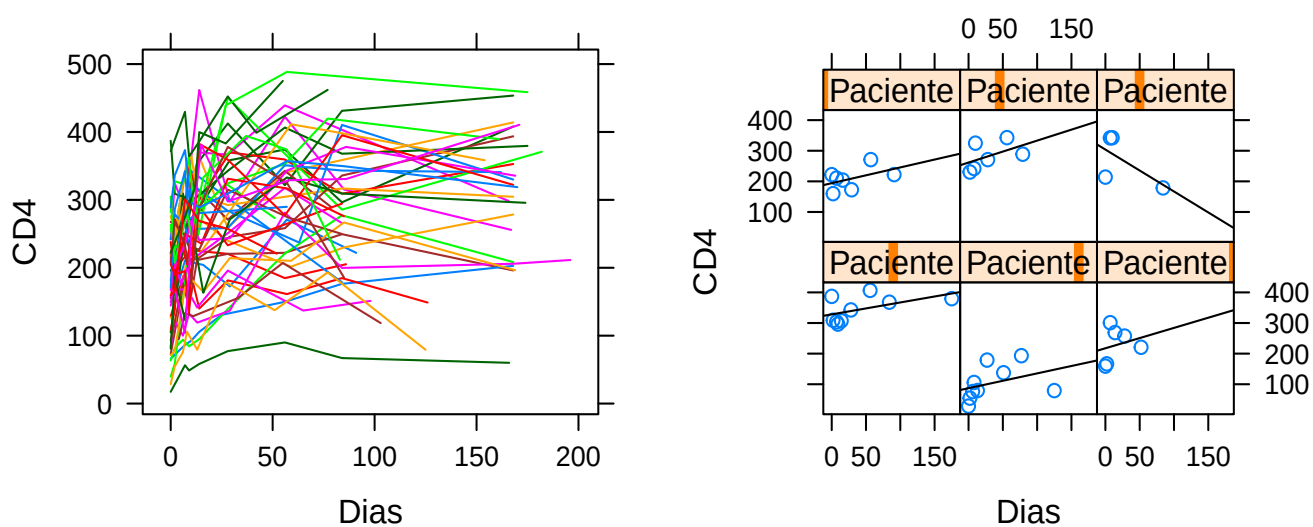


Figura 5: Gráfica en "spaghetti" de la población y tendencia lineal de algunos individuos del conteo de linfocitos CD4 a través del tiempo con los datos ACTG-315.

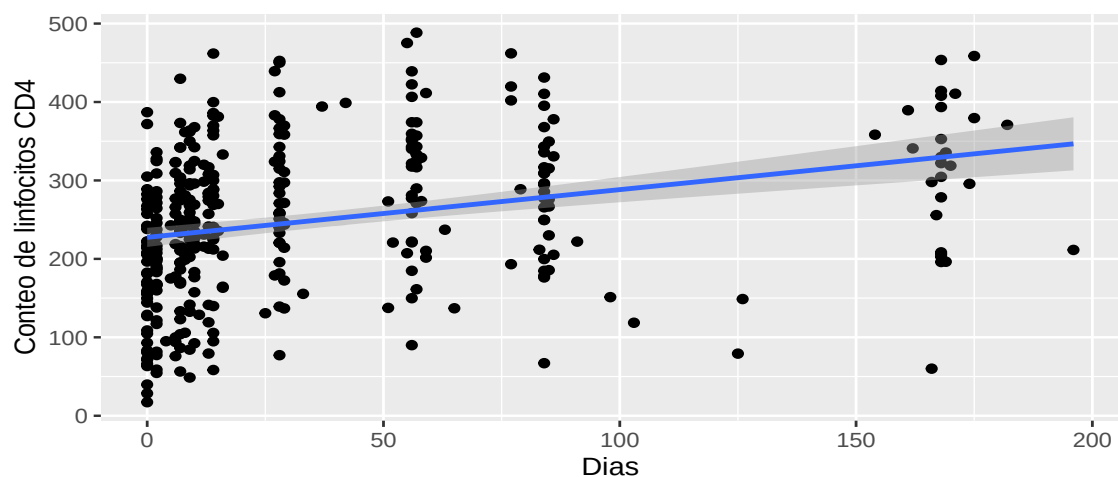


Figura 6: Tendencia lineal del conteo de linfocitos CD4 a través del tiempo en ACTG-315.

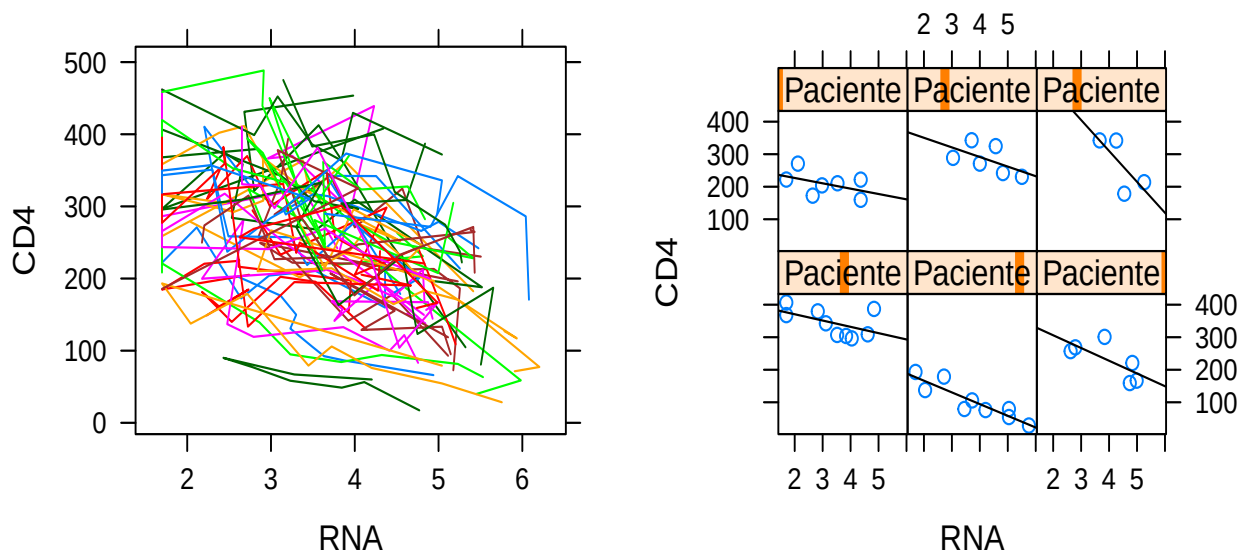


Figura 7: Gráfica en "spaghetti" de la población y tendencia lineal de algunos individuos del efecto de la carga viral sobre el conteo de linfocitos CD4 con los datos ACTG-315.

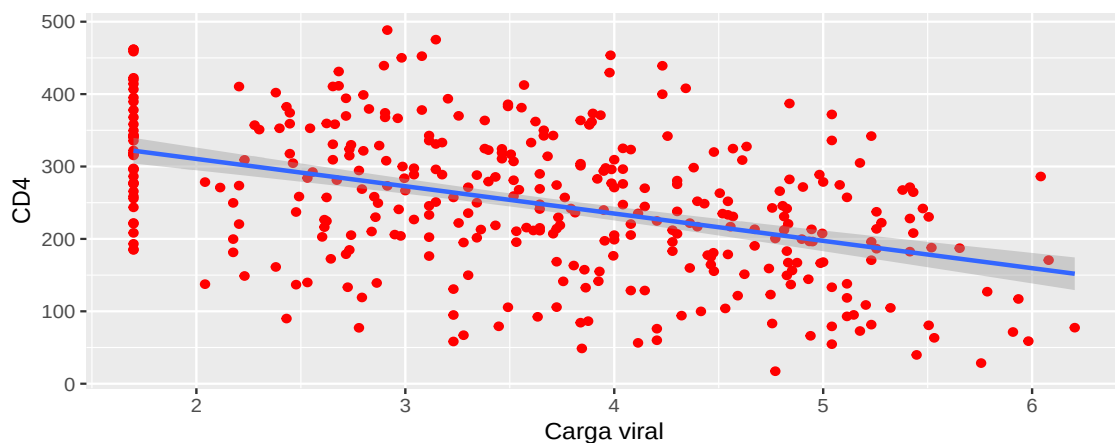


Figura 8: Tendencia lineal del conteo de linfocitos CD4 frente a la carga viral en ACTG-315.

Es importante analizar la respuesta clínica al tratamiento en diferentes momentos, lo que se puede realizar de forma paramétrica (por ejemplo con una regresión lineal), obteniéndose estimaciones segmentadas o interrumpidas de las trayectorias en cada tiempo. Esta aproximación permite observar que la relación no es constante en cada instante, ver figura 9.

Por lo tanto, como el modelado anterior no permite obtener una trayectoria continua en la relación de las variables CD4 y carga viral en cada punto del rango, se logra es un conjunto discreto de coeficientes, uno para cada tiempo.

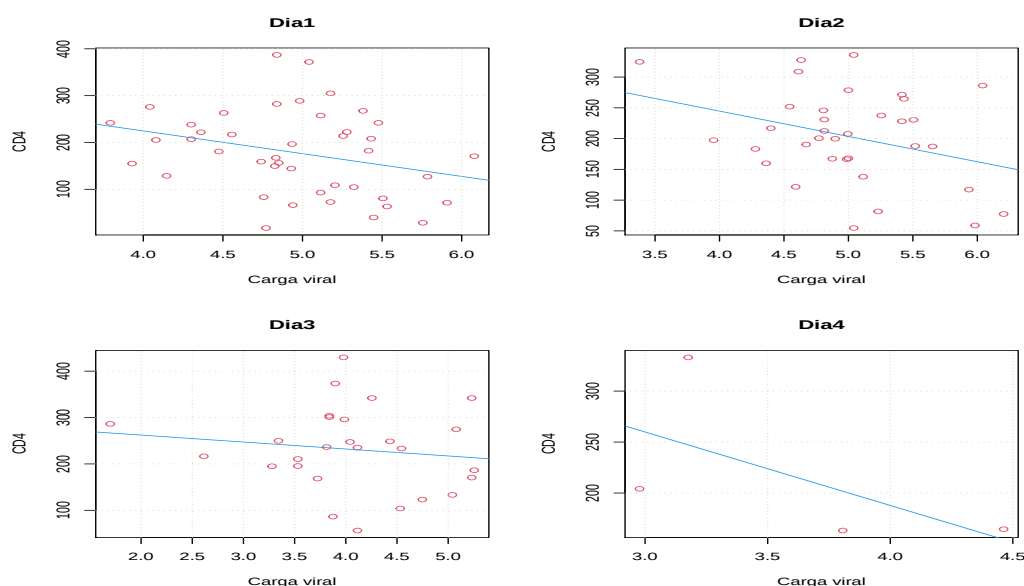


Figura 9: Tendencia del conteo de linfocitos CD4 frente a la carga viral con ACTG-315, escogiendo en algunos puntos en el tiempo de seguimiento.

Ahora bien, para obtener una curva continua de coeficientes dinámicos, el modelo no paramétrico generalizado que se propone para esta aplicación es de la forma:

$$\log(E(CD4_{ij})) = \beta_0(t_{ij}) + RNA(t_{ij})\beta_1(t_{ij}), \quad j = 1, \dots, n_i, \quad i = 1, \dots, n, \quad (4.18)$$

donde $CD4_{ij}$ es el conteo de linfocitos CD4 que se asume inicialmente una variable con distribución Poisson, $RNA(t_{ij})$ es la carga viral (en escala logarítmica) asociado con la j -ésima medida del i -ésimo paciente. $\beta_0(t_{ij})$ es el coeficiente dinámico asociado con el intercepto, y $\beta_1(t_{ij})$ es el coeficiente dinámico asociado con la carga viral. Los coeficientes asociados al intercepto y a la carga viral se pueden expresar como en 2.10 o como en 2.7, según se estimen mediante Kernel o Spline, respectivamente.

Con la data ACTG-315 se emplea el AIC como método para obtener los parámetros óptimos de suavizamiento. El modelo Spline que mejor desempeño presenta es con 3 nodos para cada coeficiente y con un polinomio de segundo grado. Con un Kernel Gaussiano, el ancho de banda seleccionado fue de 0.3181818 para el α_0 y de 0.2409091 para el de α_1 , se establecieron 5 nodos para cada coeficiente.

Para evaluar la calidad y el ajuste del modelo, al igual que con la data anterior se emplea la RMSE y el AIC. En general se observa que emplear una distribución bn es mejor que una distribución de Poisson, para modelar la variable respuesta, tanto para Kernel como para Spline. Este resultado se confirma con la prueba de Young, sin embargo esta misma prueba no distingue entre la metodología Spline y la Kernel usando una bn, ver la tabla 2. La hipótesis nula establecida para la prueba de bondad de ajuste sobre la *deviance residual* es que el modelo propuesto se ajusta adecuadamente a los datos. El resultado se presenta en la tabla 2, como $p(\chi^2)$. Para la data ACTG-315 la prueba no fue significativa, indicando que los modelos Kernel y Spline se ajustan bien a los datos, y que no hay problemas de dispersión.

Una vez estimados los α para el intercepto y la pendiente, se obtienen los $\beta(t)$. En las gráficas 10 y 11 se ilustran los resultado tanto por el método Kernel como por Spline con la data ACTG- 315. El coeficiente dinámico $\beta_1(t)$, es decir la pendiente dinámica, representa la relación dinámica del conteo de linfocitos CD4 frente a la carga viral. Los intervalos de confianza de los $\beta(t)$ para el caso de ACTG-315 por medio de bootstrap se presentan en las figuras 12 y 13.

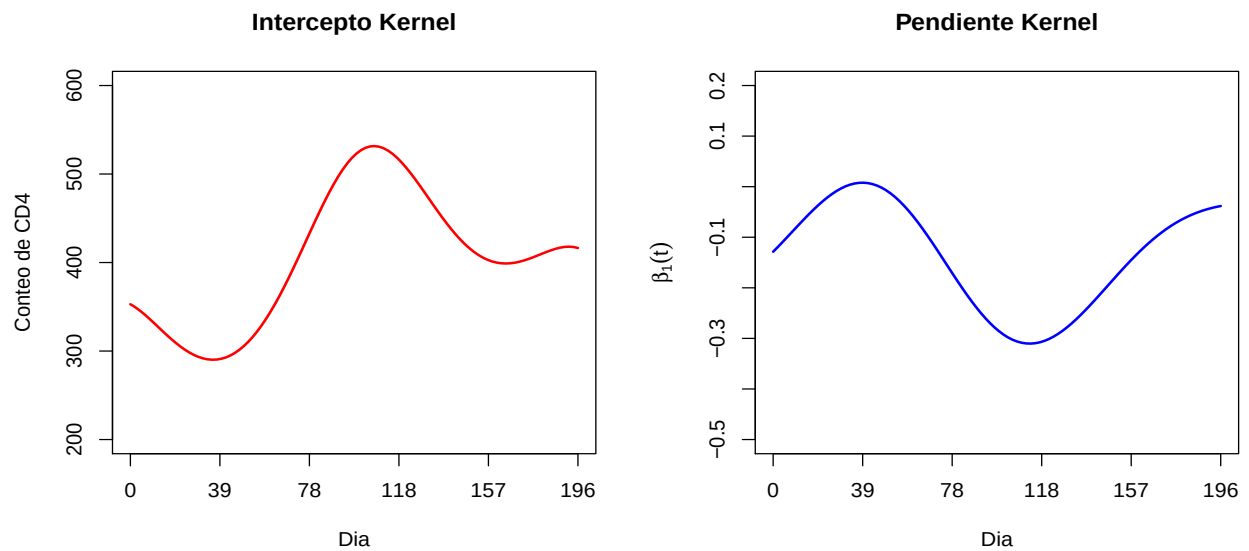


Figura 10: Intercepto (β_0) en valores originales y pendiente (β_1) del conteo de linfocitos CD4 en ACTG-315 en el tiempo con kernel.

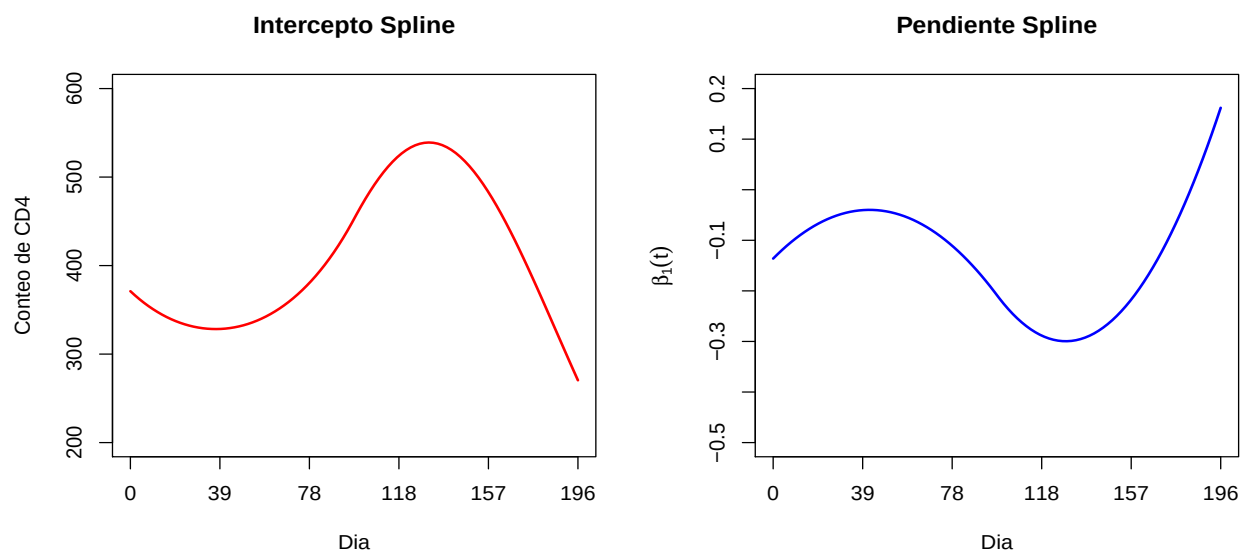


Figura 11: Intercepto (β_0) dado en valores originales y pendiente (β_1) del conteo de linfocitos CD4 en ACTG-315 en el tiempo con splines.

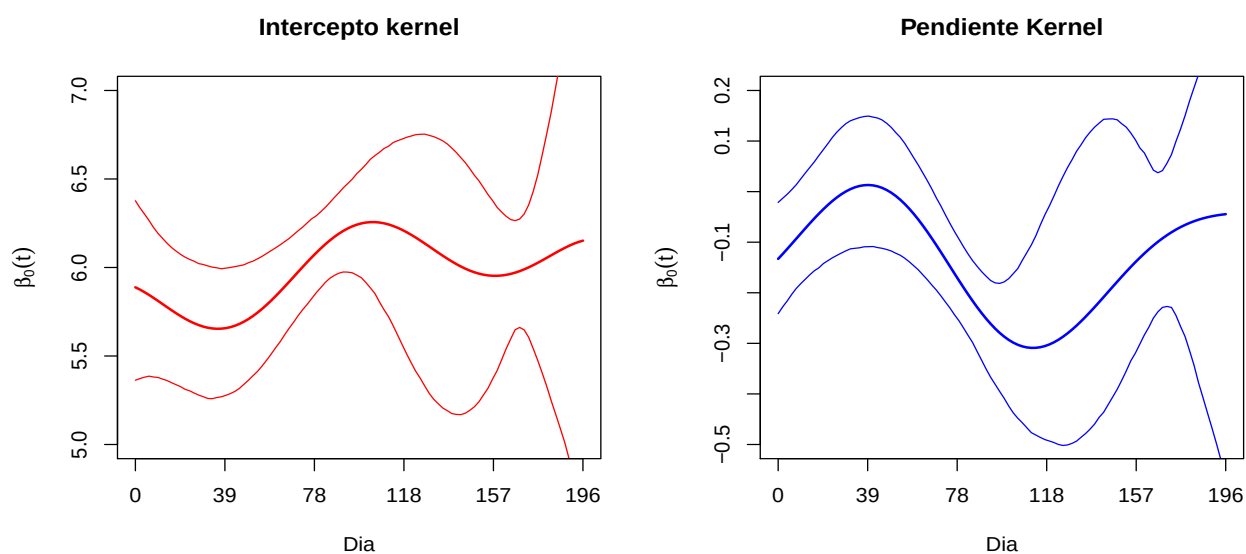


Figura 12: Modelo Kernel del contéo de linfocitos CD4 con los datos ACTG-315.

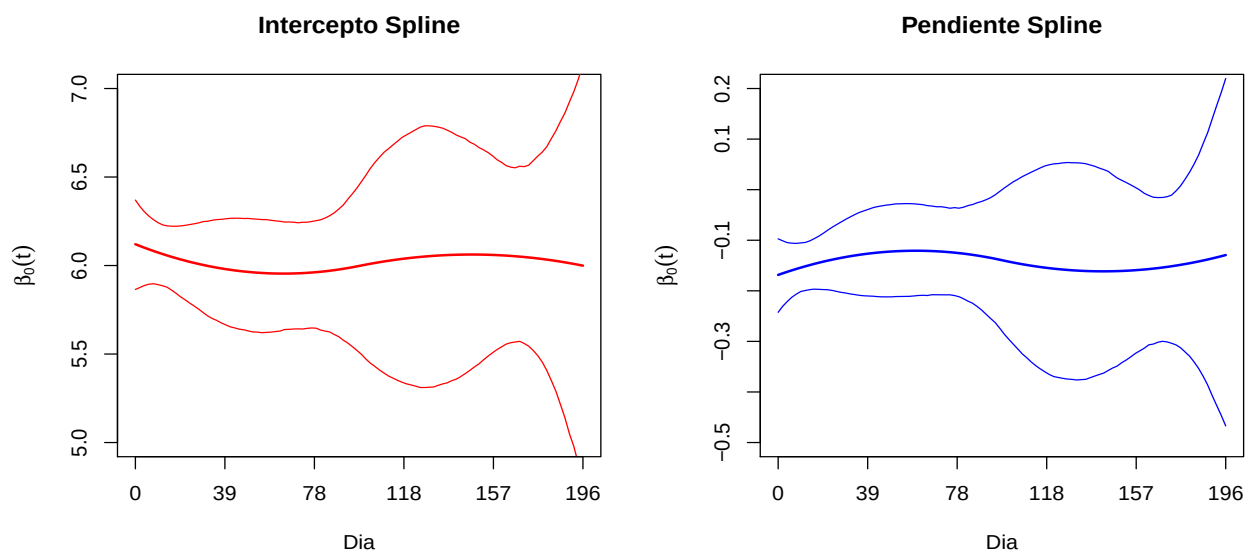


Figura 13: Modelo Spline del contéo de linfocitos CD4 con los datos ACTG-315.

Modelo	AIC	RMSE	$p(\chi^2)$
Kernel Poisson	58.8025	85.1946	1
Kernel bn	33.873	85.3114	1
Spline Poisson	55.3224	86.4135	1
Spline bn	29.882	87.4956	1

Tabla 2: Desempeño del modelo generalizado para estimar α con los datos ACTG-315.

Tanto con Spline como con Kernel se observa una correlación positiva en la pendiente, desde el comienzo hasta el día 39, para luego hacerse negativa hasta el día 118, y finalmente continuar siendo positiva hasta el final del seguimiento. Esto puede estar señalando que el tratamiento es efectivo a partir del día 39 hasta el 118, en donde el conteo de CD4 se incrementa, pero luego hay una pérdida de efectividad con disminución del conteo hasta el final. Con Kernel se observa una estabilización del conteo a partir del día 157. Con Spline se sobreesuaviza el comportamiento dinámico, mientras que con Kernel la dinámica se presenta más detallada. En ambos, las bandas al final se hacen imprecisas por pocas observaciones, existe entonces incertidumbre, cosa que al principio del tratamiento no ocurre.

En un trabajo inicial utilizando regresión spline (RS) y una base de B-splines (BSB) de segundo grado, Sosa and Díaz (2009) modelaron la relación de las variables en estudio con la data ACTG-315, encontrando que la relación al principio del tratamiento no es tan fuerte como entre los días 100 y 140 donde la relación es claramente inversa, sin embargo entre los días 140 y 150 parece que la relación es directa, pero se nota que vuelve a ser débil. En este ejercicio la variable respuesta es la carga viral y emplean una distribución normal para modelarla. En los residuales observan que el modelo ajusta bien a los datos, aunque parece haber falta de bondad de ajuste en los primeros 25 días, donde el valor de los residuales es más alto. En un trabajo ulterior del mismo autor (Sosa and Buitrago, 2021) con la misma data ACTG-315, con un seguimiento tan solo hasta el día 86, aplicando funciones radiales Kernel y regresión Spline, encontraron una relación directa hasta el día 34 luego de lo cual se pierde la relación para hacerse negativa hasta el final del periodo. Lo que puede estar indicando un inicio de efectividad del tratamiento solo hasta después del día 34. Aplicando el DIC como medida de bondad de ajuste concluyen que el TVCM basada en funciones kernel es el modelo que mejor desempeño presenta.

4.3. Aplicación caso COVID-19.

A diferencia del análisis de la relación entre el conteo de linfocitos CD4 y la carga viral en VIH/SIDA, en donde interesa es entender la dinámica poblacional e individual en relación a la respuesta al tratamiento en el horizonte temporal del estudio o rango de los datos, para el caso del COVID-19 el esfuerzo inicial está en poder predecir los casos futuros por las implicaciones en las decisiones de política sanitaria.

La forma del modelo para esta aplicación es similar al modelo 4.17, con un coeficiente $\beta_0(t_{ij})$ como intercepto, sin covariable y como variable respuesta el logaritmo del conteo de casos.

Para ejemplificar las estrategias de estimación propuestas y verificar su desempeño en las predicciones, se toma un periodo de tiempo de 58 días del 2020/marzo/6 al 2020/mayo/2, como observaciones de entrenamiento, y de 5 días para realizar las predicciones y validación. Los resultados se presentan en las figuras 14 y 15.

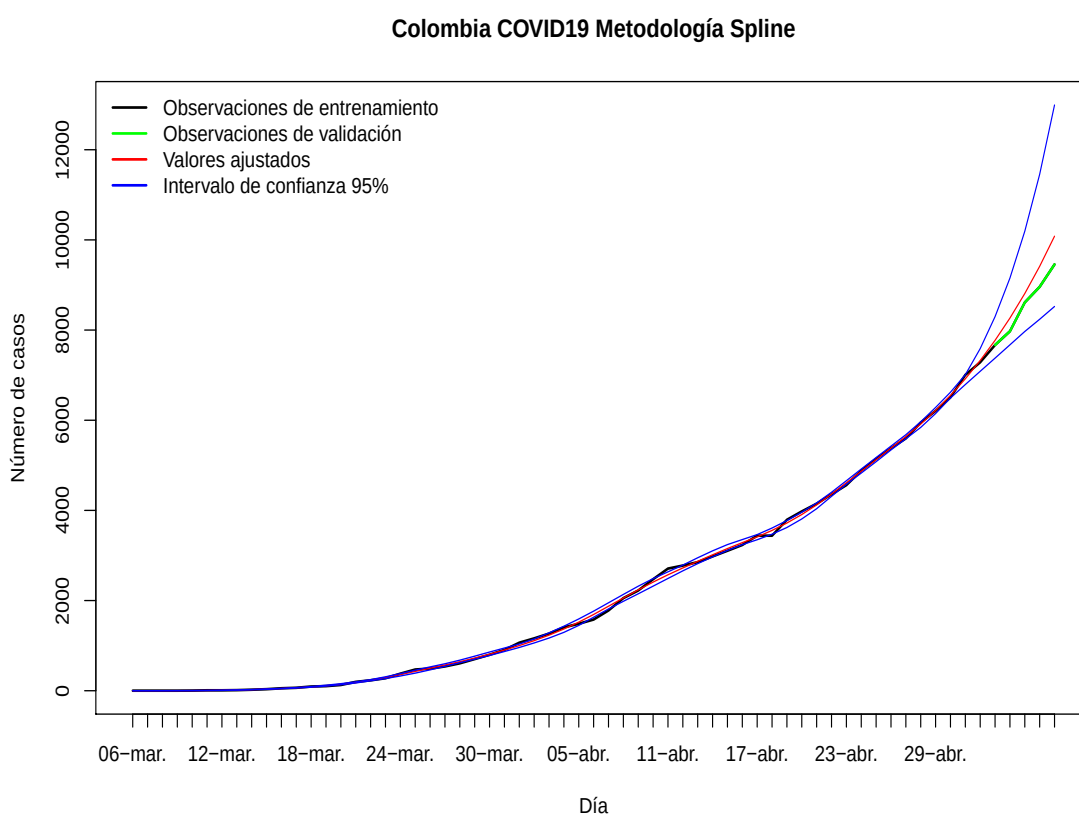


Figura 14: Modelo Spline para los datos de COVID-19 en Colombia.

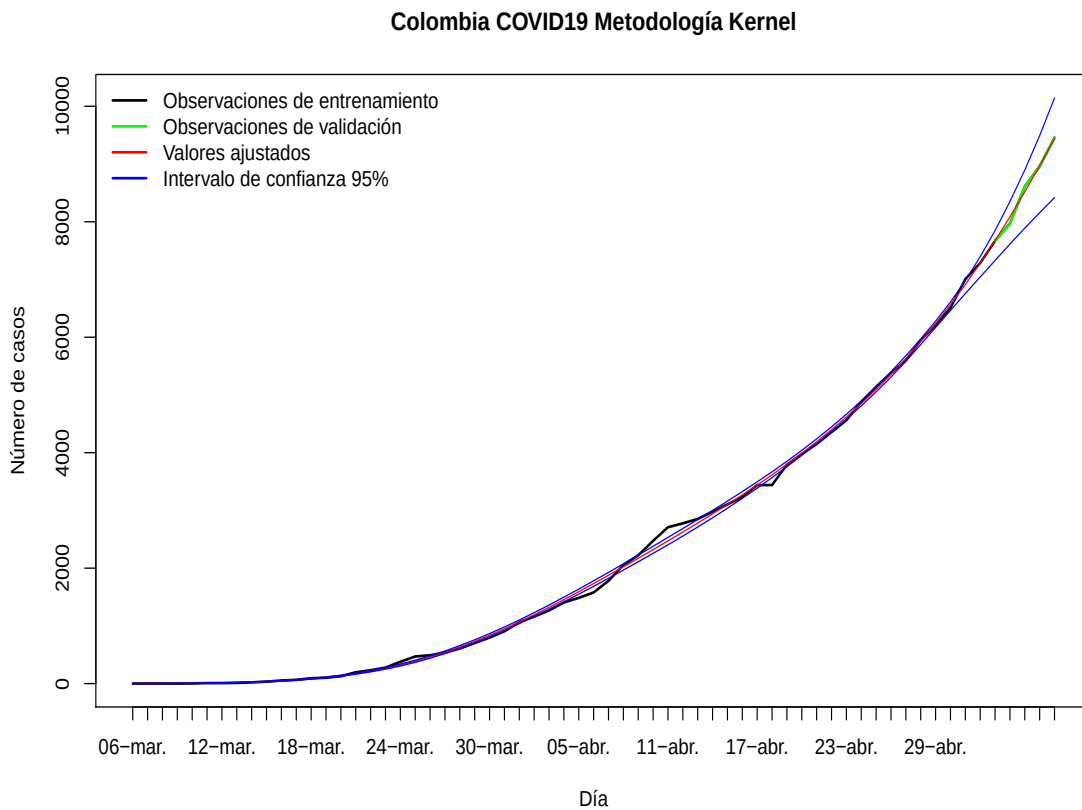


Figura 15: Modelo kernel para los datos de COVID-19 en colombia.

En la tabla 3 se presentan los resultados de la evaluación de los modelos usando distintas funciones para la variable respuesta. Los parámetros de suavizamiento se obtuvieron aplicando PCV. Con Spline se emplea un polinomio de grado 2, y 10 nodos. Con kernel se emplean 5 nodos y un ancho de banda de 0.7982983.

Modelo	AIC	RMSE	$p(\chi^2)$
Kernel Poisson	624.1957	64.05078	1.853374e-10
Kernel bn	599.2826	77.23511	0.1790205
Spline Poisson	556.6131	44.03699	0.05398627
Spline bn	558.6182	44.03909	0.05455642

Tabla 3: Desempeño de los modelos de coeficientes dinámicos generalizado para estimar α con los datos COVID-19, Colombia de marzo a mayo 2020.

Aplicando la prueba de Vounge que permite comparar los distintos modelos, se encuentra que la mejor metodología es la Spline frente a la Kernel, tanto con una función de Poisson como con una binomial negativa, y que los modelos Spline no son distinguibles en su desempeño.

5. Discusión y Conclusiones

El presente trabajo desarrolla modelos de coeficientes dinámicos generalizados para datos longitudinales con aplicaciones en salud. Dos estrategias de suavizamiento son empleadas, una con bases de potencias truncadas para spline, y otra usando un kernel Gaussiano. Se propusieron estas estrategias por una parte como alternativas a los modelos paramétricos convencionales, que estiman un número finito y constante de coeficientes, que pueden no ser la mejor opción cuando se quiere capturar la naturaleza dinámica de la variable respuesta, y por otra parte como alternativa a modelos no paramétricos que modelan la variable respuesta con una distribución normal, los cuales pueden ser problemáticos cuando los errores presentan comportamientos sesgados.

Partiendo de ciertos problemas planteados en la literatura estadística a trabajos previos (Huang et al., 2012; Lu and Huang, 2017), que modelan los datos del ensayo clínico estudiados aquí, con coeficientes dinámicos y asumiendo que la variable respuesta exhibe un comportamiento normal (Sosa and Díaz, 2009, 2012; Huang and Lu, 2016), propusimos varias soluciones alternativas para superar los problemas que se pueden presentar cuando los errores pueden exhibir comportamientos sesgados. Por una parte, se asumió que la variable respuesta es un conteo y se aplicó una distribución de Poisson para modelar el valor esperado de dicha variable; por otro lado, se empleó un modelo lineal generalizado para estimar los coeficientes dinámicos, y se usó una función de enlace conocida sobre la variable respuesta.

Los modelos de coeficientes dinámicos no paramétricos, son poco usados en salud. De hecho, en la revisión que se realizó para COVID-19, un tema muy actual, la mayoría de los modelos empleados eran paramétricos. Los modelos propuestos se ejemplificaron aplicándolos a datos que son propios del sector salud, se modeló la relación dinámica de dos variables en VIH/SIDA, lo que permitió entender el comportamiento dinámico interno de los datos en el rango de los valores analizados, útil en la monitorización de los tratamiento. En el tercer ejemplo, se predijeron casos de la actual pandemia COVID-19, ilustración útil para conocer el desempeño de los modelos propuestos por fuera del rango de los datos, al estimarse el comportamiento extrapolable de la información.

Desde el punto de vista del desempeño de los modelos desarrollados se encontró, tanto con el protocolo ACTG-388 como con la data ACTG-315, que no se distinguen las metodologías kernel y la Spline al usar en ambos una binomial negativa como función de enlace. Sin embargo, con los datos de la COVID-19 la metodología Spline resultó mejor que la Kernel, pero no es distinguible usar una función de enlace Poisson o binomial negativa al suavizar con un Spline.

En general con los datos ACTG-388, el aumento del conteo de linfocitos CD4 es mas pronunciado en las primeras 40 semanas de tratamiento y continúa incrementándose pero mas despacio entre las semanas 101 ó 112, según el método empleado, a partir del cual el conteo declina. Lo cual podría señalar un efecto incrementado del tratamiento hasta casi el final del tiempo de seguimiento, con decline en las últimas semanas. Con los datos del protocolo ACTG-315 se observa que el tratamiento se hace efectivo a partir del día 39 y hasta el día 118, a partir del cual se produce perdida de efectividad del mismo.

Los datos del protocolo ACTG empleados en el presente trabajo, han sido ampliamente estudiados en otros desarrollos de modelos con coeficientes dinámicos, lo que permitió hacer comparaciones con los estudios previos, obteniéndose resultados similares en este trabajo, lo que de alguna manera valida nuestros modelos. Sin embargo, es una limitante el no contar con datos nacionales que permitiera obtener aplicaciones inmediatas en nuestro entorno y sería de interés ampliar el análisis con mas covariables.

En el caso COVID-19, por cuestiones de alcance con los objetivos propuestos, solo se modeló el caso colombiano, pero sería muy pertinente hacerlo con otros países y regiones, contando así con datos repetidos para mas de un individuo. También se podrían predecir periodos de tiempo mas largos, para saber cuál es el máximo de tiempo que estos modelos podrían predecir y verificar si logran detectar los picos de la epidemia, así como comparar los resultados de los modelos propuestos con los paramétricos convencionales, que han sido los mas usados para describir la pandemia.

Entre las limitaciones de los modelos TVCM se tienen que solo estiman el comportamiento promedio de la población, y teniendo en cuenta que con el LDA se tienen medidas repetidas para cada individuo, es

menester realizar una extensión empleando un modelo mixto de coeficientes dinámicos y aleatorios para estimar cada trayectoria, útil para la monitorización clínica de cada paciente. Otra limitación del presente estudio es que no modeló la estructura de correlación de las medidas repetidas, que es importante porque las medidas repetidas de un mismo individuo pueden estar correlacionadas. La estrategia no paramétrica no presupone la especificación del modelo, y es especialmente útil cuando se quiere captar la dependencia no lineal de la variable respuesta a unas covariables que dependen del tiempo, sin embargo presenta debilidades cuando las covariables presentan alta dimensionalidad (Fan and Zhang, 2008).

Otras extensiones del presente trabajo podrían ser empleando otros suavizadores, y realizar estimaciones bayesianas. Hay temas en salud que se podrían beneficiar de los modelos propuestos, por ejemplo estimando las trayectorias de la resistencia antimicrobiana, específicamente en *staphylococcus aureus*, comparando países y regiones, lo que resultaría muy útil para apoyar los programas de vigilancia del uso adecuado de antibióticos.

6. Notación

Los vectores denotados por letras minúsculas en negrilla, como **a**, **b**, **x**, **y**, se asumen como vectores columna. Las letras mayúsculas en negrilla, como **A**, **B**, **X**, **Y**, denotan matrices. La transpuesta de un vector **x** se denota con $\mathbf{x}^T$, al igual que la transpuesta de una matriz **X** que se representa con $\mathbf{X}^T$. El símbolo $\equiv$ significa igual por definición.

Referencias

- Ahmadi, A., Fadaei, Y., Shirani, M., and Rahmani, F. (2020). Modeling and forecasting trend of covid-19 epidemic in iran until may 13, 2020. *Medical Journal of the Islamic Republic of Iran*, 34:27.
- Barbosa, I., de Lima, K., and de Almeida Medeiros, A. (2020). Covid-19 in brazil: analysis of the pandemic short-term scenario in relation to other countries. *Int. J. Dev. Res.*, 10(6):36840–36845.
- Batista, M. (2020). Estimation of the final size of the covid-19 epidemic. *MedRxiv*. doi, 10:16–20023606.
- Cai, Z., Fan, J., and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, 95(451):888–902.
- Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press.
- Carroll, C., Bhattacharjee, S., Chen, Y., Dubey, P., Fan, J., Gajardo, A., Zhou, X., Mueller, H.-G., and Wang, J.-L. (2020). Time dynamics of covid-19. *medRxiv*.
- Clara-Rahola, J. (2020). An empirical model for the spread and reduction of the covid19 pandemic. *Studies of Applied Economics*, 38(2).
- Cuevas, A. (2004). El análisis estadístico de grandes masas de datos: algunas tendencias recientes. *De la aritmética al análisis: historia y desarrollos recientes en matemáticas*, page 59.
- Davidian, M. (2007). *Applied Longitudinal Data Analysis*.
- de Pereda Sebastián, D., Ramos, A. M., and Ivorr, B. (2010). *Modelización matemática de la difusión de una epidemia de peste porcina entre granjas*. PhD thesis, Tesis de Maester. Universidad Complutense de Madrid. 19, 21.
- Díaz-Narváez, V., San-Martín-Roldán, D., Calzadilla-Núñez, A., San-Martín-Roldán, P., Parody-Muñoz, A., and Robledo-Veloso, G. (2020). Which curve provides the best explanation of the growth in confirmed covid-19 cases in chile? *Revista Latino-Americana de Enfermagem*, 28.

- Egger, M., May, M., Chêne, G., Phillips, A. N., Ledergerber, B., Dabis, F., Costagliola, D., Monforte, A. D., De Wolf, F., Reiss, P., et al. (2002). Prognosis of hiv-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies. *The Lancet*, 360(9327):119–129.
- Ekum, M. and Ogunsanya, A. (2020). Application of hierarchical polynomial regression models to predict transmission of covid-19 at global level. *Int. J. Clin. Biostat. Biom*, 6:027.
- Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface*, 1(1):179.
- Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press.
- Fischl, M. A., Giuliano, M., Vella, S., Ribaud, H. J., Collier, A. C., Erice, A., Dehlenger, M., Eron, Joseph J., J., Saag, M. S., Hammer, S. M., and Morse, G. D. (2003). A Randomized Trial of 2 Different 4-Drug Antiretroviral Regimens versus a 3-Drug Regimen, in Advanced Human Immunodeficiency Virus Disease. *The Journal of Infectious Diseases*, 188(5):625–634.
- Ford, N., Stinson, K., Gale, H., Mills, E. J., Stevens, W., González, M. P., Markby, J., and Hill, A. (2015). Cd4 changes among virologically suppressed patients on antiretroviral therapy: a systematic review and meta-analysis. *Journal of the International AIDS Society*, 18(1):20061.
- Gonzalo-Gil, E., Ikediobi, U., and Sutton, R. E. (2017). Mechanisms of virologic control and clinical characteristics of hiv+ elite/viremic controllers. *The Yale journal of biology and medicine*, 90(2):245.
- Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779.
- Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822.
- Huang, Y., Chen, J., and Yan, C. (2012). Mixed-effects joint models with skew-normal distribution for hiv dynamic response with missing and mismeasured time-varying covariate. *The international journal of biostatistics*, 8(1).
- Huang, Y. and Lu, T. (2016). Bayesian inference on partially linear mixed-effects joint models for longitudinal data with multiple features. *Computational Statistics*, 32.
- Hughes, M. D., Johnson, V. A., Hirsch, M. S., Bremer, J. W., Elbeik, T., Erice, A., Kuritzkes, D. R., Scott, W. A., Spector, S. A., Basgoz, N., et al. (1997). Monitoring plasma hiv-1 rna levels in addition to cd4+ lymphocyte count improves assessment of antiretroviral therapeutic response. *Annals of internal medicine*, 126(12):929–938.
- Kasilingam, D., Sathiyar Prabhakaran, S. P., Rajendran, D. K., Rajagopal, V., Santhosh Kumar, T., and Soundararaj, A. (2020). Exploring the growth of covid-19 cases using exponential modelling across 42 countries and predicting signs of early containment using machine learning. *Transboundary and Emerging Diseases*.
- Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721.
- Korenromp, E. L., Williams, B. G., Schmid, G. P., and Dye, C. (2009). Clinical prognostic value of rna viral load and cd4 cell counts during untreated hiv1 infection a quantitative review. *PloS one*, 4(6):e5950.
- Kumar, N. (2015). Review of innovation diffusion models.

- Lederman, M. M., Connick, E., Landay, A., Kuritzkes, D. R., Spritzler, J., St. Clair, M., Kotzin, B. L., Fox, L., Heath Chiozzi, M., Leonard, J. M., et al. (1998). Immunologic responses associated with 12 weeks of combination antiretroviral therapy consisting of zidovudine, lamivudine, and ritonavir: results of aids clinical trials group protocol 315. *Journal of Infectious Diseases*, 178(1):70–79.
- León Isorna, S. (2015). Modelización estadística de datos longitudinales.
- Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., et al. (2020). Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*.
- Liang, H., Wu, H., and 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.
- Lipsitz, S. R., Ibrahim, J., and Molenberghs, G. (2000). Using a box–cox transformation in the analysis of longitudinal data with incomplete responses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(3):287–296.
- Lu, T. and Huang, Y. (2017). Bayesian inference on mixed-effects varying-coefficient joint models with skew-t distribution for longitudinal data with multiple features. *Statistical methods in medical research*, 26(3):1146–1164.
- Lu, Y. and Zhang, R. (2009). Smoothing spline estimation of generalised varying-coefficient mixed model. *Journal of Nonparametric Statistics*, 21(7):815–825.
- Luo, J. (2020). Predictive monitoring of covid-19. *SUTD Data-Driven Innovation Lab*.
- Ma, J. (2020). Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling*, 5:129–141.
- Manrique Abril, F., González-Chordá, V. M., Gutiérrez Lesmes, O. A., Tellez Piñerez, C. F., Herrera-Amaya, G. M., et al. (2020). Modelo sir de la pandemia de covid-19 en colombia.
- Marschner, I. C., Collier, A. C., Coombs, R. W., D’aquila, R. T., DeGruttola, V., Fischl, M. A., Hammer, S. M., Hughes, M. D., Johnson, V. A., Katzenstein, D. A., et al. (1998). Use of changes in plasma levels of human immunodeficiency virus type 1 rna to assess the clinical benefit of antiretroviral therapy. *Journal of Infectious Diseases*, 177(1):40–47.
- Mercker, M., Betzin, U., and Wilken, D. (2020). What influences covid-19 infection rates: A statistical approach to identify promising factors applied to infection data from germany. *medRxiv*.
- Millán-Oñate, J., Rodríguez-Morales, A. J., Camacho-Moreno, G., Mendoza-Ramírez, H., Rodríguez-Sabogal, I. A., and Álvarez-Moreno, C. (2020). A new emerging zoonotic virus of concern: the 2019 novel coronavirus (sars cov-2). *Infectio*, 24(3):187–192.
- Ministerio (2014). Guía de práctica clínica basada en la evidencia científica para la atención de la infección por vih/sida en adolescentes (con 13 años de edad o más) y adultos.
- Montgomery, D., Peck, E., and Vining, G. G. (2006). Introducción al análisis de regresión lineal. *México: Limusa Wiley*.
- Park, T. and Jeong, S. (2018). Analysis of poisson varying-coefficient models with autoregression. *Statistics*, 52(1):34–49.
- Pelinovsky, E., Kurkin, A., Kurkina, O., Kokoulina, M., and Epifanova, A. (2020). Logistic equation and covid-19. *Chaos, Solitons & Fractals*, 140:110241.
- Ramsay, J., Hooker, G., and Graves, S. (2009). Functional data analysis with r and matlab: Springer science & business media.

- Şentürk, D. and Müller, H.-G. (2008). Generalized varying coefficient models for longitudinal data. *Biometrika*, 95(3):653–666.
- Sosa, J. and Buitrago, L. (2021). Time-varying coefficient model estimation through radial basis functions. *arXiv preprint arXiv:2103.00315*.
- Sosa, J. and Díaz, L. (2009). *Desarrollo de un modelo de coeficientes dinámicos y aleatorios para el análisis longitudinales*. PhD thesis, Tesis de maestría, Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia.
- Sosa, J. C. and Díaz, L. G. (2012). Random time-varying coefficient model estimation through radial basis functions. *Revista Colombiana de Estadística*, 35(1):167–184.
- Thiébaud, R., Morlat, P., Jacqmin-Gadda, H., Neau, D., Mercié, P., Dabis, F., Chêne, G., et al. (2000). Clinical progression of hiv-1 infection according to the viral response during the first year of antiretroviral treatment. *Aids*, 14(8):971–978.
- Trujillo, C. H. S. (2020). Resumen: Consenso colombiano de atención, diagnóstico y manejo de la infección por sars-cov-2/covid-19 en establecimientos de atención de la salud. *Infectio*, 24(3).
- Twisk, J. W. (2013). *Applied longitudinal data analysis for epidemiology: a practical guide*. Cambridge University Press.
- Wang, P., Zheng, X., Li, J., and Zhu, B. (2020). Prediction of epidemic trends in covid-19 with logistic model and machine learning technics. *Chaos, Solitons & Fractals*, 139:110058.
- Wang, Y. (2007). Varying-coefficient models: New models, inference procedures, and applications.
- WHO (2007). Who case definitions of hiv for surveillance and revised clinical staging and immunological classification of hiv-related disease in adults and children.
- WHO et al. (2016). *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach*. World Health Organization.
- Wu, C. O. and Tian, X. (2018). *Nonparametric Models for Longitudinal Data: With Implementation in R*. Chapman and Hall/CRC.
- Wu, H. and Zhang, J.-T. (2006). *Nonparametric regression methods for longitudinal data analysis: mixed-effects modeling approaches*, volume 515. John Wiley & Sons.
- Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, pages 689–699.