UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Julio 2025, Vol. 18, No. 1, pp. 103–136

Generalized Dynamic Coefficient Models for Longitudinal Data Analysis in Health: An Application in HIV/AIDS and COVID-19

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

Juan Sosa[a]
jcsosam@unal.edu.co

Elkin Javier Casadiego Rincón[b]
elkincasadiego@usantotomas.edu.co

Abstract

Longitudinal data analysis is essential when the response variable is measured repeatedly on the same observational unit over time. Traditionally, parametric methods have been employed to estimate coefficients that define the relationship between the linear predictor and the response variable. However, these methods fail when the assumptions regarding the response variable and the model's random components are violated, or when the expected value of the response variable (or its transformation via a link function) cannot be expressed as a known function of the fixed and random effects. In such cases, parametric models may yield conclusions that deviate significantly from the dataset's average trends. Non-parametric regression techniques, which replace fixed parameters with time-dependent smoothed local functions, offer a powerful alternative for longitudinal data analysis. These methods, known as dynamic coefficients or parameters, allow us to establish a more flexible functional relationship between the response variable and the covariates. In this study, we propose estimation and inference techniques for generalized non-parametric dynamic coefficient models, particularly for count response variables. We illustrate our approach through its application in analyzing the effect of viral load on CD4 cell count in HIV/AIDS patients undergoing antiretroviral therapy, as well as in predicting COVID-19 cases.

***Keywords*:** Longitudinal data analysis, radial basis kernel functions, regression splines, time-varying coefficient models, viral load, CD4 T lymphocyte counts, HIV/AIDS, and COVID-19..

Resumen

[a]Departamento de Estadística, Universidad Nacional de Colombia
[b]Facultad de Estadística, Universidad Santo Tomas

El análisis de datos longitudinales es esencial cuando la variable de respuesta se mide repetidamente en la misma unidad de observación a lo largo del tiempo. Tradicionalmente, se han empleado métodos paramétricos para estimar los coeficientes que definen la relación entre el predictor lineal y la variable de respuesta. Sin embargo, estos métodos fallan cuando no se cumplen las suposiciones sobre la variable de respuesta y los componentes aleatorios del modelo, o cuando el valor esperado de la variable de respuesta (o su transformación mediante una función de enlace) no puede expresarse como una función conocida de los efectos fijos y aleatorios. En tales casos, los modelos paramétricos pueden generar conclusiones que se desvían significativamente de las tendencias promedio del conjunto de datos. Las técnicas de regresión no paramétrica, que reemplazan los parámetros fijos por funciones locales suavizadas dependientes del tiempo, ofrecen una alternativa poderosa para el análisis de datos longitudinales. Estos métodos, conocidos como coeficientes o parámetros dinámicos, nos permiten establecer una relación funcional más flexible entre la variable de respuesta y las covariables. En este estudio, proponemos técnicas de estimación e inferencia para modelos generalizados no paramétricos de coeficientes dinámicos, particularmente para variables de respuesta de conteo. Ilustramos nuestro enfoque mediante su aplicación al análisis del efecto de la carga viral sobre el conteo de células CD4 en pacientes con VIH/SIDA en tratamiento antirretroviral, así como en la predicción de casos de COVID-19.

***Palabras clave*:** Análisis de datos longitudinales, función kernel de base radial, spline de regresión, modelo de coeficientes variables en el tiempo, carga viral, conteo de linfocitos T CD4, VIH/SIDA, COVID-19..

1. Introduction

Longitudinal data analysis (LDA) is relevant when a set of n individuals is repeatedly observed over time, recording the values of a response variable along with the corresponding covariates, which may or may not depend on the time at which they are measured. In such cases, observations within the same subject are not independent, whereas measurements across different individuals are considered independent (Twisk, 2013). The response variable can be modeled by estimating either the population-level average trend or the trajectory of each individual, the latter being characterized as a mixed-effects or random-effects model.

The relationship between the response variable and the covariates in longitudinal data analysis (LDA) can be modeled using a parametric approach by fully specifying the distribution of the response variable through a fixed and finite set of coefficients, naturally extending the linear regression model for cross-sectional data (Davidian et al., 2008). In parametric models, coefficients have immediate interpretability, and commonly used estimation techniques, such as ordinary/generalized least squares, maximum likelihood, and restricted maximum likelihood, are directly applicable (Cuevas, 2004). Two notable techniques are generalized estimating equations (GEE) and linear mixed-effects models (LMEM). In GEE, the response variable is not required to follow a normal distribution, and the possible correlation among repeated measurements within a subject is explicitly accounted for using a working correlation matrix. In contrast, LMEM explicitly models within-subject variation by fitting a normal distribution around specific regression parameters and estimating the variance of this distribution, allowing for direct modeling of individual-specific characteristics (Twisk 2013; see Chapter 4 for more details).

The potential disadvantages of parametric models arise when the assumptions regarding the response variable and the random component of the model are not met, or when the response variable (or a function of this variable through a link function) does not correspond to a known function of the fixed and random effects. These issues can lead parametric models to produce conclusions that deviate from the average trend in the dataset. In such cases, non-parametric regression techniques offer a highly robust modeling alternative. One type of non-parametric model that facilitates linear discriminant analysis (LDA) is the time-varying coefficient model (TVCM), where local smoothed functions that vary with time, referred to as dynamic coefficients or parameters, replace fixed parameters. These models allow for a more flexible functional relationship between the response variable and the covariates (Sosa and Díaz, 2012; Lu and Huang, 2017; Wu and Tian, 2018). The non-parametric approach does not require model specification *a priori*, making it particularly useful for capturing the linear dependence of a response variable on time-dependent covariates. However, it faces challenges when the covariates exhibit high dimensionality (Fan and Zhang, 2008).

A common approach to modeling the mean response in a TVCM is to assume that random errors follow a Gaussian process. However, several authors, including Huang et al. (2012) and Lu and Huang (2017), have highlighted a limitation in earlier works (e.g., Sosa and Díaz 2009, Sosa and Díaz 2012, Huang and Lu 2016), where a symmetric distribution (such as the normal distribution) was assumed for the response variable, despite potential skewed behavior in the errors. Consequently, statistical inference based on this normality assumption may lead to inaccurate results.

In some cases, issues with skewness in the errors can be addressed by applying transformations to the response variable, such as linearizing the data through a logarithmic scale. When the response function appears curvilinear, transformations can help maintain the model as first-order. A lower-order model for a transformed variable is often preferable to a higher-order model using the original metric (Montgomery et al., 2006). However, if a low-order polynomial provides a poor fit to the data and increasing the polynomial order does not significantly improve the fit, alternative solutions to transforming the response variable should be considered. This issue may arise when the function behaves differently across various segments of the covariate range. A well-known solution for handling skewed response variables involves applying the Box-Cox transformation, which enables the transformed variable to have a conditional Gaussian distribution, as demonstrated in an LDA by Lipsitz et al. (2000). One limitation of this approach is that the transformation form is invariant over time. For an LDA, the response variable may exhibit different skewness patterns over time, making a fixed transformation unsuitable for achieving a Gaussian approximation across the entire range of interest (Wu and Tian, 2018). An effective way to address this limitation is to account for the dynamic nature of the variable using a generalized model and incorporating non-parametric adjustments for the covariates, as proposed in this study.

Consequently, an extension of TVCM for a non-normal response variable is a generalized model that introduces a known link function to relate the dynamic linear predictor and the response variable. In this context, several developments have been proposed. Hastie and Tibshirani (1993) introduced varying-coefficient models where covariates may or may not depend on time, coefficients are smoothed functions, and the mean response is linked to the linear predictor through a link function; the model was exemplified for binary responses. Wu and Zhang (2006) developed generalized dynamic coefficient models with a logistic link function for longitudinal binary response data, referring to this approach as the generalized nonparametric population mean (GNPM) model. Fan and Zhang (2008) proposed statistical methods for generalized varying-coefficient models, including those designed for longitudinal data. Lu and Zhang (2009) introduced a generalized varying-coefficient mixed model using dynamic coefficients to estimate the mean function. Şentürk and Müller (2008) developed a generalized dynamic coefficient model for longitudinal data, accounting for lag effects in the linear predictor. Park and Jeong (2018) presented a Poisson autoregressive varying-coefficient model, incorporating an autocorrelation structure and illustrating it with daily homicide data. See also Wang (2007) for additional examples.

Despite the advancements in generalized non-parametric dynamic coefficient models, previous works by Sosa and Díaz (2009); Sosa and Díaz (2012) and Sosa and Buitrago (2022) did not account for potential asymmetries in error behavior within the analyzed applications. To address these limitations, this study proposes novel solutions. First, the response variable is assumed to be a count, and a Poisson distribution is applied to model its expected value. Additionally, a generalized linear model is employed to estimate the dynamic coefficients, incorporating a known link function for the response variable.

The models developed in this study are exemplified by analyzing the dynamic relationship between two variables in the context of HIV/AIDS and the prediction of COVID-19 case numbers. The first example underscores the importance of using TVCMs (time-varying coefficient models) to monitor patient responses to treatment in a clinical setting. The second example demonstrates the methodology's applicability in an epidemiological context, particularly for forecasting case numbers, which is crucial for informed decision-making in public health.

The data used in the illustrations for the HIV/AIDS case studies come from two clinical trials conducted by the AIDS Clinical Trials Group (ACTG). The first study, ACTG-315, included 46 patients infected with HIV-1 who were treated with an antiretroviral therapy regimen consisting of ritonavir, zidovudine, and lamivudine. Viral load and CD4 counts were measured simultaneously on days 0, 2, 7, 10, 14, 28, 56, 84, 168, and 196 (Lederman et al., 1998). The second study, ACTG-388, involved 517 patients with advanced disease who were treated with indinavir plus either efavirenz or nelfinavir. Further details about this dataset are provided in the applications section (Fischl et al., 2003). The COVID-19 case study in Colombia utilizes data obtained from official repositories hosted on government websites. For more information about the AIDS Clinical Trials Group, visit `https://actgnetwork.org/`. Furthermore, COVID-19 data can be accessed via the official repository: `https://www.datos.gov.co/api/views/gt2j-8ykr/rows.csv?accessType=DOWNLOAD`.

The objective of this study is to develop estimation techniques for dynamic coefficients in count data using spline regression and radial kernel functions (RKF) within the framework of a generalized linear model. These techniques aim to characterize the evolution of the effect of viral load on CD4 cell counts in HIV/AIDS patients undergoing antiretroviral therapy. The methodology is also exemplified by forecasting the number of cases over a specified time period during the COVID-19 pandemic in Colombia. Additionally, inference methods for the model coefficients are developed using resampling techniques. Finally, the performance of the estimation is assessed in terms of precision and efficiency through goodness-of-fit measures and error metrics.

The document is organized as follows: Section 2 presents the theoretical components of linear dynamic coefficient models. Also, this section develops the generalized non-parametric model employed in the applications, including smoothing strategies using splines and kernel functions. Section 3 provides the context of the health-related examples used to illustrate the models. Finally, the last two sections present the results, discussion, and conclusions.

2. Dynamic Coefficient Models

In this section we introduce dynamic coefficient models, including linear and generalized versions, to analyze longitudinal data with time-varying relationships. It details estimation techniques like spline regression and kernel functions, addressing non-normal responses and flexible modeling of time-varying effects.

2.1. Linear Dynamic Coefficient Model

A semiparametric model for LDA in the context of CD4 cell count applications was proposed by Zeger and Diggle (1994). A generalization of this model, incorporating time-varying coefficients (TVCM), was later developed by Hoover et al. (1998). The model is given by:

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

where n is the number of individuals in the study, n_i is the number of measurements for the i-th individual, t_{ij} represents the time at which the j-th measurement for the i-th individual was taken, and y_{ij} denotes the response variable associated with the i-th individual at time t_{ij}. The term $\boldsymbol{x}_{ij} = \boldsymbol{x}_i(t_{ij}) = [x_{i0}(t_{ij}), \ldots, x_{id}(t_{ij})]^\top$ is a $(d+1) \times 1$ vector of covariates measured at time t_{ij} for the i-th individual, whose components may or may not vary with time.

The vector $\boldsymbol{\beta}(t_{ij}) = [\beta_0(t_{ij}), \beta_1(t_{ij}), \ldots, \beta_d(t_{ij})]^\top$ represents the $(d+1)\times 1$ dynamic coefficients (also referred to as dynamic parameters) evaluated at time t_{ij}. The error term $\epsilon_{ij} = \epsilon_{ij}(t_{ij})$ denotes the measurement error for the i-th individual at time t_{ij}, assumed to be a stochastic process with zero mean, independent of $\boldsymbol{x}_{ij}$, and with covariance function $\Gamma(t, t') = \mathsf{Cov}[\epsilon(t), \epsilon(t')]$. Additionally, measurements across individuals are assumed to be independent (Şentürk and Müller, 2008). Initially, there are no strong restrictions on the $(d+1)$ functions comprising the dynamic coefficient vector, other than requiring them to be sufficiently smooth (differentiable) to enable the application of estimation procedures (Sosa and Díaz, 2009).

When model (2.1) excludes covariates, i.e., $d = 0$ and $x_{i0}(t_{ij}) = 1$ for all $i = 1, \ldots, n$, it can be expressed as:

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

Model (2.2) will be used in the ACTG-388 application, as this dataset does not include covariates, and also for the COVID-19 application. If model (2.1) considers only one time-dependent covariate, i.e., $d = 1$ and $x_{i0}(t_{ij}) = 1$ for all $i = 1, \ldots, n$, it can be expressed as:

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

This model (2.3) will be applied to the ACTG-315 dataset, which includes the covariate viral load.

In a TVCM, the response variable can be either continuous quantitative or discrete. In the latter case, it may be modeled using a Gaussian process when the discrete distribution is symmetric. The model presented thus far is designed to handle quantitative variables assumed to follow a normal distribution, which is not always appropriate for reasons discussed in the introduction. Therefore, it is necessary to extend the model to account for the nature of other types of variables, such as count data.

2.2. Generalized Linear Dynamic Coefficient Model

An extension of a TVCM for a non-normal response variable is a generalized model that introduces a known link function to relate the linear dynamic predictor and the response variable. In a generalized linear model for LDA, the marginal population mean and variance of the responses y_{ij} are specified as:

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

where ϕ is a scale parameter, w_{ij} are weights, and $V(\cdot)$ is a known variance function that represents the dependence of the variance on the mean response. In a generalized dynamic coefficient model, the marginal mean μ_{ij} is assumed to be related to the linear predictor $\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}(t_{ij})$ through a link function $g(\cdot)$. The model is expressed as (Cai et al. 2000, Wu and Zhang 2006, Wang 2007):

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}(t_{ij}), \quad j = 1, 2, \ldots, n_i,\ i = 1, 2, \ldots, n. \tag{2.4}$$

In our experiments, the weight for each individual is considered as $w_i = 1/(nn_i)$, for $i = 1, \ldots, n$, where each subject is inversely weighted by their number of repeated measures n_i. This ensures that individuals with fewer repeated measures receive more weight than those with more measurements. Two non-parametric estimation methods for the functions $\beta_r(t)$, $r = 0, 1, \ldots, d$, are described below, based on basis functions (Wu and Zhang 2006 and Ramsay et al. 2009): a global estimation using regression splines (RS) and a local estimation using radial kernel functions (RKF; Wu and Tian 2018).

2.2.1. Estimation Using Spline Regression

Spline regression is a method that expresses each dynamic coefficient $\beta(t)$ in model (2.4) as a linear combination of spline basis functions (Wu and Tian, 2018). A spline basis $\phi(t)$ represents the individual terms or blocks comprising the linear combination of basis functions (also referred to as a basis function expansion; Ramsay et al. 2009). In other words, spline regression can be viewed as a piecewise polynomial of order $g+1$ or degree g, where $g \geq 0$. The degree refers to the highest power in the polynomial (Wu and Zhang, 2006).

Specifically, a piecewise polynomial is a real-valued function $f(\cdot)$ defined over the range of interest, an interval $[a, b]$, which is divided into a number of partitions:

$$a = \tau_0 < \tau_1 < \tau_2 < \ldots < \tau_K < \tau_{K+1} = b.$$

These partitions form contiguous subintervals, and $\tau_0, \tau_2, \ldots, \tau_K$ are referred to as knots, where K is the number of knots (Wu and Zhang, 2006). Various techniques are used to represent spline basis functions, including polynomial bases, M-splines, I-splines, truncated power basis (TPB), and B-spline basis (BSB).

A truncated power basis (TPB) is a set of spline basis functions that can be represented (Wu and Zhang, 2006) with g degrees and K knots as follows:

$$1,\ t,\ \ldots,\ t^g,\ (t-\tau_1)_+^g,\ \ldots,\ (t-\tau_K)_+^g, \tag{2.5}$$

where $(t-\tau_k)_+^g$ denotes the g-th power of the positive part of $(t-\tau_k)$, defined as $(t-\tau_k)_+^g = \text{máx}(0, t-\tau_k)^g$.

In summary, the set (2.5) of $K+g+1$ basis functions is known as the truncated power basis of order $g+1$ or degree g, with interior knots τ_ℓ, $\ell = 1, 2, \ldots, K$.

Using the set (2.5), it is possible to express a spline regression, meaning each dynamic coefficient $\beta_r(t)$, $r = 0, 1, \ldots, d$, as:

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

where $\boldsymbol{\alpha}_r = (\alpha_{r,0}, \alpha_{r,1}, \ldots, \alpha_{r,K+g+1})^\top$ is a $p_r \times 1$ column vector of unknown coefficients, and $p_r = K + g + 1$ is the number of basis functions.

Now, for convenience, the TPB (2.5) can be represented as:

$$\boldsymbol{\Phi}_{r,p_r}(t) = [1, t, \ldots, t^g, (t-\tau_1)_+^g, \ldots, (t-\tau_K)_+^g]^\top,$$

where $\boldsymbol{\Phi}_{r,p_r}(t)$ denotes the basis functions evaluated at time t, and p_r is a non-negative integer indicating the number of basis functions. Using this representation, the spline regression $\beta_r(t)$ in (2.6) can be rewritten as:

$$\beta_r(t) = \boldsymbol{\Phi}_{r,p_r}(t)^\top \boldsymbol{\alpha}_r. \tag{2.7}$$

Substituting $\beta_r(t)$ from (2.7) into (2.4), the generalized dynamic coefficient model, including covariates, becomes:

$$g(\mu_{ij}) = \boldsymbol{z}_{ij}^\top \boldsymbol{\alpha}, \quad j = 1, 2, \ldots, n_i,\ i = 1, 2, \ldots, n, \tag{2.8}$$

where $y_{ij} \equiv y_i(t_{ij})$, and $\boldsymbol{z}_{ij} = (\boldsymbol{z}_{0,ij}^\top, \ldots, \boldsymbol{z}_{d,ij}^\top)^\top$, with:

$$\boldsymbol{z}_{r,ij} = x_{r,i}(t_{ij}) \boldsymbol{\Phi}_r(t_{ij}), \quad r = 0, 1, \ldots, d.$$

Similar to $\boldsymbol{\alpha}$, each $\boldsymbol{z}_{ij}$ is a column vector, representing the covariate multiplied by the vector of basis functions of dimension $p \times 1$. For the i-th subject, for $i = 1, 2, \ldots, n$, the response vector and design matrix are denoted as:

$$\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,n_i})^\top, \quad \mathbf{Z}_i = [\boldsymbol{z}_{i,1}, \ldots, \boldsymbol{z}_{i,n_i}]^\top.$$

Consequently, the response vector and design matrix for all the data are expressed as:

$$\boldsymbol{y} = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_n^\top)^\top, \quad \mathbf{Z} = [\mathbf{Z}_1^\top, \ldots, \mathbf{Z}_n^\top].$$

This allows (2.8) to be written in matrix form as a standard generalized linear model:

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

where $\boldsymbol{\mu}$ is the expected value of the response vector $\boldsymbol{y}$, i.e., $\boldsymbol{\mu} = \mathsf{E}(\boldsymbol{y})$. Once $\boldsymbol{\alpha}$ is estimated using an appropriate method, an estimate of $\beta(t)$ can be obtained.

2.2.2. Estimation Using Kernels as Basis Functions

Kernels are functions associated with each data point, where the weighted sum of these functions serves as an estimator to approximate an unknown density function. The definition of the kernel estimator is determined by two parameters: the bandwidth h and the kernel function $k(\cdot)$ (Wu and Tian, 2018).

The kernel function K defines the shape of the weights assigned to each observation within the defined bandwidth, thereby determining their importance in the estimation process. Common choices for the kernel function in radial kernel functions

(RKF) include:

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

The bandwidth h is a positive parameter that initially corresponds to the range of the estimation interval. However, selecting an appropriate bandwidth is a critical aspect of the estimation procedure, as the resulting estimator's properties heavily depend on this choice. There exists a trade-off between flexibility and complexity controlled by the bandwidth parameter. If h is chosen too small, only observations very close to the estimation point will contribute, leading to a highly variable estimated curve that captures local behaviors well but lacks smoothness. Conversely, if h is too large, the estimates at each point will be influenced by observations far from the estimation point, resulting in a smoother curve that misses local patterns and introduces significant bias. From a theoretical perspective, the bandwidth selection should strike a balance between bias and variance to achieve an optimal trade-off in the estimation process.

2.2.3. Estimation Using RKF as Basis Functions

The estimation strategy using RKF involves approximating each component of $\boldsymbol{\beta}(t)$ using RKF, treating them as radial basis functions (RBF). This approach models $\beta_r(t),\ r = 0, 1, \ldots, d$, through the expression:

$$\beta_r(t) = \sum_{\ell=1}^{M} \xi_r \left(\frac{|t - t_\ell|}{h}\right) \alpha_{r\ell} = \boldsymbol{\Xi}_r(t)^\top \boldsymbol{\alpha}_r, \tag{2.10}$$

where M is the number of distinct measurement times t_ℓ, $|\cdot|$ denotes the absolute value, ξ_r is a kernel function, h is the bandwidth, and $\boldsymbol{\alpha}_r = [\alpha_{r1}, \ldots, \alpha_{rM}]^\top$ is the $M \times 1$ vector of associated coefficients. $\boldsymbol{\Xi}_r(t)$ is an $M \times 1$ vector comprising the kernel functions used for approximation:

$$\boldsymbol{\Xi}_r(t) = \left[\xi_r\left(\frac{|t - t_1|}{h}\right), \ldots, \xi_r\left(\frac{|t - t_M|}{h}\right)\right]^\top. \tag{2.11}$$

Substituting $\beta_r(t)$ from (2.10) into (2.4), the generalized dynamic coefficient model, including covariates, becomes:

$$g(\mu_{ij}) = \boldsymbol{v}_{ij}^\top \boldsymbol{\alpha}, \quad j = 1, 2, \ldots, n_i;\ i = 1, 2, \ldots, n, \tag{2.12}$$

where $y_{ij} \equiv y_i(t_{ij})$, and $\boldsymbol{v}_{ij} = (\boldsymbol{v}_{0,ij}^\top, \ldots, \boldsymbol{v}_{d,ij}^\top)^\top$, with:

$$\boldsymbol{v}_{r,ij} = x_{r,i}(t_{ij})\boldsymbol{\Xi}_r(t_{ij}), \quad r = 0, 1, \ldots, d.$$

Similar to $\boldsymbol{\alpha}$, each $\boldsymbol{v}_{ij}$ is a column vector representing the covariate multiplied by the vector of basis functions of dimension $p\times 1$. For the i-th subject ($i = 1, 2, \ldots, n$), the response vector and design matrix are denoted as:

$$\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,n_i})^\top, \quad \mathbf{V}_i = [\boldsymbol{v}_{i,1}, \ldots, \boldsymbol{v}_{i,n_i}]^\top.$$

Consequently, the response vector and design matrix for all the data are expressed as:

$$\boldsymbol{y} = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_n^\top)^\top, \quad \mathbf{V} = [\mathbf{V}_1^\top, \ldots, \mathbf{V}_n^\top].$$

This allows (2.12) to be written in matrix form as a standard generalized linear model:

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

where $\boldsymbol{\mu}$ is the expected value of the response vector $\boldsymbol{y}$, i.e., $\boldsymbol{\mu} = \mathsf{E}(\boldsymbol{y})$. Similar to spline regression, once $\boldsymbol{\alpha}$ is estimated using an appropriate method, $\boldsymbol{\beta}(t)$ can be directly obtained.

The procedure used to estimate the coefficients $\boldsymbol{\alpha}$ in (2.9) or (2.13), once the optimal knots and bandwidth are determined, is based on maximum likelihood estimation via iteratively reweighted least squares (IRLS). This is implemented by default in the `glm.fit` function from the `glm` package in `R`.

2.2.4. Estimation of Smoothing Parameters

The selection of the number of knots and bandwidth (collectively referred to as smoothing parameters, $\boldsymbol{\rho}$) is a critical step in the estimation process. For spline regression, the smoothing parameter vector corresponds to the selection of knots, $\boldsymbol{\rho} = [p_0, \ldots, p_d]^\top$, whereas for RKF, it also includes bandwidth selection, $\boldsymbol{\rho} = [h_0, \ldots, h_d]^\top$. The model's fit depends significantly on the optimal selection of these parameters.

Smoothing parameters are directly related to the set of basis functions used to express the dynamic coefficients. For truncated power bases (TPB), the smoothing parameters are given by:

$$\rho_r = K_r + g_r + 1, \quad r = 0, 1, \ldots, d, \tag{2.14}$$

where K_r is the number of knots and g_r is the degree of the polynomial basis associated with the estimation of the r-th dynamic coefficient. Selecting the smoothing parameters involves determining K_r and g_r. Typically, the polynomial degree is chosen as linear ($g = 1$), quadratic ($g = 2$), or cubic ($g = 3$), with cubic polynomials being preferred in most references. The number of knots, however, requires an established criterion for selection.

Knot Placement

In spline regression, once the number of knots K_r is determined using an appropriate criterion, they must be placed across the range of interest. For longitudinal data, this typically involves evenly spacing the knots along the temporal interval $[a, b]$, defined by the minimum and maximum measurement times:

$$a = \text{mín}(t_{ij} : i = 1, \dots, n,\ j = 1, \dots, n_i),\ b = \text{máx}(t_{ij} : i = 1, \dots, n,\ j = 1, \dots, n_i).$$

In practice, measurement times are scaled so that the minimum value is 0 and the maximum is 1. This scaling facilitates domain segmentation and smooth parameter tuning (e.g., knot number and bandwidth).

Selection of Number of Knots and Bandwidth

For normal models, the optimal number of knots and bandwidth is often determined through leave-one-out cross-validation (LOOCV) (Wu and Zhang, 2006; Hoover et al., 1998). LOOCV minimizes the following expression:

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

where $w_i = 1/(nn_i)$ is the weight for the i-th individual, and $\hat{\boldsymbol{\beta}}_{(-ij)}(t_{ij})$ is an estimate of $\boldsymbol{\beta}(t_{ij})$ obtained excluding the j-th measurement of the i-th individual.

Let $\mathbf{A}$ be the smoothing matrix defined as $\mathbf{A} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$, where $\mathbf{X}$ is the basis function matrix (spline or kernel). The fitted values are given by $\hat{\boldsymbol{y}} = \mathbf{A}\boldsymbol{y}$. Wu and Zhang (2006) showed that (2.15) is equivalent to:

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

where $\boldsymbol{y}$ is the response vector, $N = \sum_{i=1}^{n} n_i$ is the total number of observations, $\mathbf{W} = \text{diag}(w_1 \mathbf{I}_{n_1}, \dots, w_n \mathbf{I}_{n_n})$ is the weighting matrix, and $\text{tr}(\mathbf{A})$ denotes the trace of $\mathbf{A}$.

While LOOCV is robust to normality deviations, we also used the Akaike Information Criterion (AIC) to select optimal smoothing parameters. AIC balances model fit and complexity and is defined as:

$$\text{AIC} = 2p - 2\ln(L),$$

where p is the number of model parameters and L is the maximum likelihood value. The best model minimizes AIC. An algorithm was constructed to select optimal parameters for β_0 and β_1 independently.

2.2.5. Bootstrap Statistical Inference

Confidence intervals for a point estimate of a parameter can be established using the bootstrap resampling technique. In the nonparametric case, where we have an estimation function for the coefficients, a confidence band is developed. Assuming independence across individuals, resampling is performed on all repeated measurements for each subject with replacement from the original data set (Wu and Tian, 2018). The following steps can be followed to obtain the estimation function by bootstrap resampling:

1. **Sampling**: Randomly select n subjects with replacement from the original sample, maintaining the same sample size.
2. **Bootstraping**: The bootstrap sample obtained through resampling is denoted by:
$$\{(Y_{ij}^*, t_{ij}^*) : i = 1, \dots, n; \, j = 1, \dots, n_i\},$$
where some values from the original sample may appear more than once in the new bootstrap sample.
3. **Design matrix**: Obtain the design matrix by applying the nonparametric smoother to the bootstrap sample obtained.
4. **Estimation**: Estimate the coefficients $\boldsymbol{\beta}_r = [\beta_{r0}, \beta_{r2}, \dots, \beta_{rK_r+g+1}]^\top$ for spline or $\boldsymbol{\beta}_r = [\beta_{r1}, \dots, \beta_{rM}]^\top$ for kernel, using the information obtained in the previous steps by applying the `glm.fit` function from the `glm` package in `R`.
5. **Nonparametric estimation**: Calculate the nonparametric estimator curve based on the bootstrap sample, denoted as $\hat{\mu}_k^{\text{boot}}(t; h; w)$ for $\mu(t)$.
6. **Repeat**: Repeat the above steps B times. The B-bootstrap estimator is denoted as $\hat{\mu}_k^{\text{boot}}(t; h; w)$, and the estimators obtained from the B samples are:
$$\hat{\mu}_B(t, h, w) = \hat{\mu}_{1k}^{\text{boot}}(t, h, w), \dots, \hat{\mu}_{Bk}^{\text{boot}}(t, h, w).$$
7. **Inference**: The generated bootstrap samples are used to approximate confidence intervals or bands for $\mu(t)$, based on the percentiles of the bootstrap samples, which are given by:
$$(L_{\alpha/2}(t), U_{1-\alpha/2}(t)),$$
where $L_{\alpha/2}(t)$ and $U_{1-\alpha/2}(t)$ are the lower and upper $\alpha/2$ and $1 - \alpha/2$ percentiles of $\hat{\mu}_B(t; h; w)$, respectively.

Nonparametric time-varying coefficient models can be computationally intensive, especially when applied to large datasets or fine-grained temporal resolutions. In this study, the sample sizes were moderate, and estimation was performed efficiently using matrix-based implementations and built-in routines such as `glm.fit` in R. However, in large-scale applications, computational demands may become a limiting factor. To address this, we can use strategies to improve efficiency, such as using low-rank basis approximations, implementing stochastic optimization techniques (e.g., subsampling or mini-batch estimation), or parallelizing bootstrap procedures. Incorporating these techniques would facilitate the application of dynamic models to high-dimensional or streaming data contexts.

3. Clinical and Epidemiological Context of the Cases Under Study

Here, we provide the clinical and epidemiological background for HIV/AIDS and COVID-19. For HIV/AIDS, it emphasizes the use of viral load and CD4 T cell count as key markers to monitor disease progression and treatment response, noting their variability and the need for dynamic modeling. For COVID-19, it highlights the role of epidemic modeling to predict case trends and support public health decisions, reviewing both classical and empirical approaches. The section underscores the importance of longitudinal data analysis in capturing dynamic relationships for both diseases.

3.1. Context of HIV/AIDS

The human immunodeficiency virus type 1 (HIV-1) is a retrovirus belonging to the family *Retroviridae*. It contains its genetic information in the form of ribonucleic acid (RNA) and causes an infection that, in most untreated individuals, leads to a decline in CD4 T lymphocytes (CD4 T cells), resulting in morbidity and mortality (Gonzalo-Gil et al., 2017). The progressive deterioration of the immune system due to HIV-1 infection culminates in advanced stages with the development of acquired immunodeficiency syndrome (AIDS). AIDS is defined by the presence of one or more than 20 opportunistic infections or cancers related to HIV (Organization, 2007).

Two key prognostic markers commonly used to monitor the progression of HIV/AIDS are quantitative HIV RNA (viral load) and CD4 T cell count. These markers play a crucial role in initial assessment and evaluating treatment responses (Ministerio de Salud y Protección Social, 2014). Viral load is particularly useful for determining prognosis, establishing a baseline for assessing treatment response, and monitoring viral suppression (Hughes et al., 1997). Regular measurement of viral load has been associated with significant clinical outcomes, including resistance assessment, management of opportunistic infections, mortality prediction, and monitoring disease progression (Marschner et al., 1998; Thiébaut et al., 2000). Higher initial viral replication levels, measured by RNA copies per milliliter (or their logarithmic transformation), predict faster progression to AIDS and death. Furthermore, clinical progression is generally preceded by an increase in viral RNA levels (Korenromp et al., 2009).

As HIV infection progresses, a decline in CD4 T cell count is observed. This decline serves as an immunological parameter that helps determine the risk of HIV-associated complications, the need for prophylaxis against opportunistic infections, and the initiation of antiretroviral treatment (Ministerio de Salud y Protección Social, 2014; Gonzalo-Gil et al., 2017). At the start of active antiretroviral therapy, CD4 T cell count has been identified as the most critical predictive factor for clinical progression. For instance, patients initiating treatment with fewer than 200 cells/μL had a significantly higher risk of clinical progression compared to those with higher counts (Egger et al., 2002).

Understanding the relationship between viral load and CD4 T cell count, their predictive value, and their role in the natural history of the infection is essential. At a population level, these markers establish thresholds for initiating treatments and prophylaxis efficiently, while at an individual level, they enable monitoring of treatment response and prediction of disease progression (Korenromp et al., 2009). However, the prognostic value of these markers varies depending on factors such as disease stage, age, patient condition (e.g., pregnancy, comorbidities, baseline immunological status), the direct impact of the virus independent of CD4 T cell effects, unknown time of infection since seroconversion, and the duration of follow-up in studies (Korenromp et al., 2009).

CD4 T cells are well-known targets of HIV and decline with disease progression. However, when antiretroviral therapy effectively suppresses viral load, CD4 T cell counts can recover to normal levels (Lu and Huang, 2017). Despite the broad consensus on the predictive nature of viral load and CD4 T cell count, their relationship is dynamic, changing with or without antiretroviral treatment. This variability necessitates the development of appropriate models to characterize their temporal relationship (Liang et al., 2003). The variability in CD4 T cell count and viral load, both within and between individuals over time, the multivariate nature of their association, and differing analytical methodologies complicate joint modeling for risk or prognostic assessment (Korenromp et al., 2009).

Generally, virological response (measured through viral load) and immunological response (measured through CD4 T cell count) are negatively correlated during antiviral treatment (Ford et al., 2015). However, this relationship can be influenced by various factors, making it non-constant during treatment. Some patients fail to achieve expected viral suppression, necessitating a deeper understanding of these dynamics to optimize treatment monitoring. Studies have confirmed that viral load and CD4 T cell count exhibit distinct response patterns, with potential discrepancies between virological and immunological responses requiring both population-level and individual-level evaluations (Liang et al., 2003; Korenromp et al., 2009).

Given these complexities, longitudinal analysis of these markers is more appropriate than cross-sectional analysis. Longitudinal strategies allow for a dynamic understanding of their relationship over time, accounting for key factors influencing the analysis. Cross-sectional regression assumes that response variable measurements are independent, while longitudinal data typically involve dependent observations from the same individual over time, warranting specialized models.

For modeling the population-level relationship between viral load and CD4 T cell count, Sosa and Díaz (2009) proposed a time-varying coefficient model (TVCM) with random errors following a Gaussian process. However, as highlighted by Huang et al. (2012) and Lu and Huang (2017), a limitation of this approach is its assumption of symmetric distributions, such as normality, for the response variable. This assumption is inappropriate for skewed error behaviors, potentially leading to misleading statistical inferences. Unlike previous studies, this work focuses on CD4 T cell count as the response variable, reflecting its critical role as a primary HIV target. This choice necessitates a generalized nonparametric dynamic model for count data.

3.2. Context of the COVID-19 Pandemic

Coronavirus Disease 2019 (COVID-19) is caused by infection with SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2), an RNA virus belonging to the subfamily *Orthocoronavirinae*, genus *Betacoronavirus* (Millán-Oñate et al., 2020; Trujillo, 2020). Coronaviruses (CoVs) periodically emerge worldwide, causing Acute Respiratory Infections (ARIs). SARS-CoV-2 was classified as a Public Health Emergency of International Concern (PHEIC) by the World Health Organization (WHO).

COVID-19 was first identified in December 2019 among patients potentially exposed to zoonotic transmission in a market in Wuhan, Hubei Province, China, and was formally recognized in January 2020 (Li et al., 2020). WHO began reporting cases on January 21, 2020, and the epidemic spread globally, with the first case confirmed in Colombia on March 6, 2020. Infection occurs when individuals are exposed to virus particles expelled by symptomatic or asymptomatic infected persons (Trujillo, 2020).

Modeling the course of the epidemic is critical for assessing its population-level impact and evaluating mitigation measures. COVID-19 dynamics have been studied through both deterministic and stochastic epidemiological models. Deterministic models group individuals into predefined states and focus on the variation of these states, represented mathematically by derivatives (Manrique Abril et al., 2020; de Pereda Sebastián et al., 2010). A foundational example is the SIR model (Susceptible-Infected-Recovered), which uses differential equations to describe population flows between states (Luo, 2020).

Epidemics are often modeled as initially exhibiting exponential growth, shaped over time by the natural course of the disease and intervention measures such as lockdowns and vaccination (Ma, 2020; Kasilingam et al., 2020). Growth curves, typically sigmoidal, capture this dynamic by reflecting a rise over time toward a saturation level (Kumar, 2015). Logistic models, a classic tool in population dynamics, are frequently used for this purpose. For instance, Kasilingam et al. (2020) used logistic regression and machine learning to predict containment signs across 42 countries, while Ma (2020) compared SIR, exponential, logistic, and Richards models. Other approaches, such as generalized logistic equations (Pelinovsky et al., 2020) and Weibull-based models (Batista, 2020), have been applied to predict epidemic trends.

Empirical models, including smoothed regression methods, offer an alternative to parametric models by relying exclusively on observed data. For example, Ekum and Ogunsanya (2020) employed polynomial regression to forecast COVID-19 cases, while Barbosa et al. (2020) used Joinpoint regression to estimate cumulative incidence across countries. Díaz-Narváez et al. (2020) explored various curve-fitting methods to describe epidemiological trends, identifying double exponential smoothing as the best fit. Clara-Rahola (2020) employed functional principal component analysis (FPCA) to predict short-term case trajectories, and Mercker et al. (2020) used generalized additive mixed models (GAMM) with autoregressive structures to explain epidemic behavior in Germany.

Building on these approaches, this study will apply generalized dynamic coefficient models to analyze COVID-19 dynamics in Colombia, providing a data-driven framework for short-term epidemic prediction.

4. Applications

In this section, we present applications of generalized non-parametric dynamic coefficient models (TVCMs) in three case studies: the ACTG-388 and ACTG-315 studies on HIV/AIDS, and COVID-19 case prediction in Colombia. In ACTG-388, TVCMs analyze CD4 count trajectories over time, revealing variability in treatment responses among patients. The ACTG-315 study examines the dynamic relationship between viral load and CD4 counts, showcasing the model's ability to estimate time-varying clinical relationships. For COVID-19, the model forecasts case trends using public health data, demonstrating its adaptability for short-term predictions and public health planning.

5. 4.1. ACTG-388 Case Study

The AIDS Clinical Trials Group (ACTG), established in 1987, is one of the world's largest HIV clinical trial networks. This example is based on ACTG protocol 388, a randomized, open-label trial involving 517 patients with advanced HIV/AIDS. Participants were eligible if their CD4 cell count was $\leq$ 200 cells/μL or their plasma HIV-1 RNA levels were $\geq 80,000$ copies/mL. Patients were assigned to three groups receiving highly active antiretroviral therapy: the reference group (indinavir plus lamivudine and zidovudine), and two comparison groups (indinavir combined with efavirenz or nelfinavir). They were monitored over 2.1 years (Fischl et al., 2003).

This analysis focuses on data from one treatment group in the ACTG-388 protocol, comprising 166 patients monitored for 120 weeks, during which CD4_{ij} cell counts were regularly measured. The dataset lacks additional covariates and illustrates the longitudinal behavior of CD4_{ij} counts. Missing observations due to non-adherence result in unbalanced data, with 1 to 18 measurements per patient, totaling 2,107 observations, and CD4_{ij} counts ranging from 0 to 1,364 cells/μL.

An initial exploration using conventional methods highlights the limitations of such approaches for longitudinal data. Figure 1 presents "spaghetti plots" showing individual CD4_{ij} trajectories over time, revealing no clear population trend. Examining a subset of individuals reveals varying patterns—some increasing, some decreasing, and others with no discernible trend. A linear regression applied to the dataset (not shown here for brevity) indicates a general upward trend in CD4_{ij} counts over time. However, this method lacks the precision needed to capture the dynamic behavior of CD4_{ij} counts at specific time points, which is crucial for monitoring treatment responses and guiding clinical decisions.

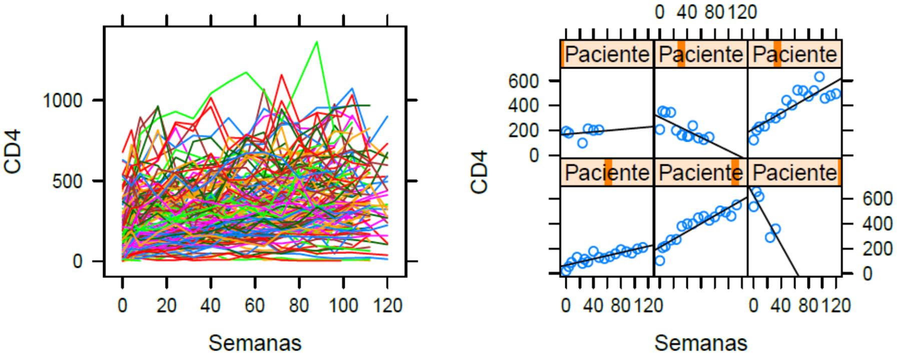


Figure 1: Spaghetti plot of the population and the linear trend of selected individuals' CD4 lymphocyte counts over time using data from ACTG-388.

To address this, a generalized non-parametric time-varying coefficient model (TVCM) was applied using the form:

$$\log \mathsf{E}(\text{CD4}_{ij}) = \beta_0(t_{ij}) \quad j = 1, \ldots, n_i,\ i = 1, \ldots, n,$$

where CD4_{ij} represents the CD4 count, assumed to follow a Poisson distribution, for the j-th measurement of the i-th patient, and $\beta_0(t_{ij})$ is the dynamic parameter for the intercept, estimated using either kernel or spline methods as described earlier.

Optimal smoothing parameters were selected via the PCV criterion (Section 2.2.3). For spline estimation, six nodes were used; for kernel estimation, six nodes and a bandwidth of 0.29 yielded the best model. After determining the optimal parameters, coefficients were estimated using the iteratively reweighted least squares (IRLS) method, implemented through the `glm.fit` function in R's `glm` package.

To evaluate the quality of the proposed models, we employed various strategies. First, we used residual deviance to indirectly assess model adequacy in estimating the coefficients (α) while also verifying for dispersion issues. The Akaike Information Criterion (AIC), which balances model fit and complexity, was utilized to select the best-performing model between the spline and kernel strategies. Additionally, we applied the root mean square error (RMSE) as a direct evaluation of the goodness of fit of the estimated responses. RMSE measures the distance between predicted and observed values and was computed using the `rmse` function from the `metrics` package in R. Models with the lowest AIC and RMSE values were selected.

We performed a goodness-of-fit test on the residual deviance using the χ^2 distribution as the decision criterion. For an adequately specified model without dispersion issues, the residual deviance should be approximately equal to its degrees of freedom. This forms the null hypothesis that the residual deviance follows a χ^2 distribution. A non-significant test result indicates no evidence of model misfit or dispersion problems (Faraway, 2016). The results showed no overdispersion in the methods, as evidenced by the χ^2 test on residual deviance, supporting our decision not to reject the null hypothesis. The goodness-of-fit based on residual deviance is presented as $p(\chi^2)$ in Table 1.

To estimate the coefficients β, we tested various distributions for modeling the response variable in the generalized linear model. The negative binomial distribution outperformed others, as indicated by the lowest AIC values, for both kernel and spline methodologies. Additionally, we applied the Vuong test to compare models under the null hypothesis that they are indistinguishable (Cameron and Trivedi, 2013). The test results confirmed our initial interpretation that the negative binomial is a better choice than the Poisson distribution. However, no significant performance difference was observed between the kernel and spline methodologies when both employed the negative binomial. Table 1 summarizes the evaluation results for model quality and fit.

Model	**AIC**	**RMSE**	$\mathbf{p(\chi^2)}$
Kernel Poisson	151.313	199.7203	1
Kernel NB	27.17015	199.9836	1
Spline Poisson	156.1622	199.4498	1
Spline NB	33.16135	199.5138	1

Table 1: Performance of the generalized model for estimating β using the ACTG-388 data.

The $\beta_0(t)$ estimates for the ACTG-388 data obtained using both methods are presented in Figure 2. To construct the confidence bands around the estimation function of the dynamic coefficients, we performed a bootstrap resampling procedure, drawing samples with replacement from the original dataset. This process generated 1,000 bootstrap samples, which we used to estimate the β coefficients, calculate the central value of $\beta_0(t)$, and determine the upper and lower confidence limits (2.5 % and 97.5 % percentiles, respectively), as described in the section on bootstrap inference. The 95 % credible confidence bands for the intercept, obtained using both the Spline and Kernel methods, are shown in Figure 3.

The results from both models exhibit similar patterns, with some notable differences. Using the Spline method, the mean trajectory of the CD4 count shows an initially steep increase, peaking around week 15 (289 cells/μL). This is followed by a continuous decline until week 26 (262 cells/μL), after which the trajectory resumes an upward trend with fluctuations until week 101 (404 cells/μL). From

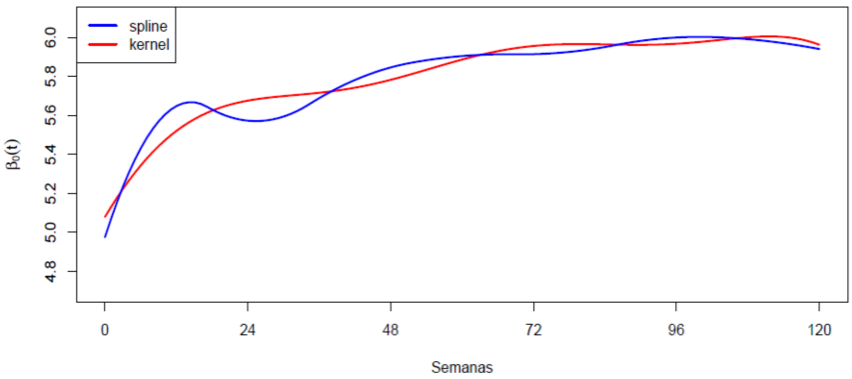


Figure 2: Estimated intercept β_0 of the CD4 lymphocyte count using ACTG-388 data.

this point, a gradual decline is observed toward the end of the period. By contrast, the Kernel method produces a smoother trajectory, with a steady increase up to week 112 (404 cells/μL), followed by a gradual decrease thereafter. Overall, the most significant growth occurs during the first 40 weeks of treatment, with a slower increase observed between weeks 101 and 112, depending on the method, after which the CD4 count begins to decline.

In summary, these results suggest that the treatment effect persists for nearly the entire follow-up period, with a slight decline in the final weeks. Similar findings were reported by Wu and Zhang (2006), who used splines with a second-degree polynomial and four nodes. Their analysis showed a rapid increase in the population mean function during the initial weeks, followed by a slower increase until approximately week 110, after which a decline was observed toward the end of the study. However, their residual analysis revealed asymmetry and significant bias in the standardized residuals, leading them to recommend a data transformation (Wu and Zhang, 2006).

In a more recent study, Sosa and Buitrago (2022) modeled the ACTG-388 data and reported similar trends. They found that the mean CD4 cell count increased rapidly during the first 40 weeks of treatment, continued to rise slowly until week 100, and then declined toward the end of the study period. Their residual analysis indicated that the model fit the data well. Using the Deviance Information Criterion (DIC) as a measure of goodness of fit, they concluded that the TVCM based on kernel functions provided the best overall performance.

5.1. ACTG-315 Case Study

The ACTG-315 protocol evaluates the impact of intensive antiretroviral therapy on immune response. It consists of a 12-week regimen combining Zidovudine, Lamivudine, and Ritonavir. The treatment was tested on 46 adults aged 18 years or

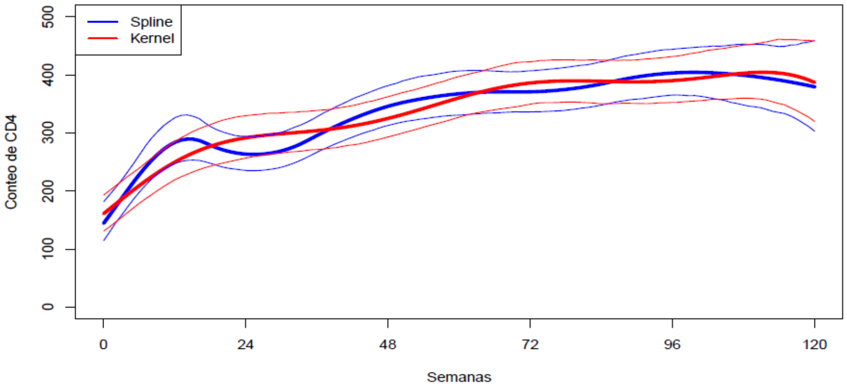


Figure 3: Confidence bands for the intercept of the CD4 lymphocyte count over the follow-up period using ACTG-388 data, comparing kernels and splines.

older with moderate HIV-1 infection (defined as CD4 lymphocyte counts between 100 and 300 cells/μL), of whom 44 completed at least nine weeks of the 12-week study duration (Lederman et al., 1998).

The dataset analyzed includes a maximum follow-up period of 196 days, with each patient contributing a varying number of repeated measurements (ranging from 4 to 10), resulting in 361 total observations. Due to this variability, the dataset is unbalanced. Viral load is reported on a logarithmic scale, while CD4 lymphocyte counts range from 17 to 488 cells/μL.

Initially, we employed conventional methodologies to describe the data, followed by analyses using the proposed generalized dynamic coefficient models. Figure 4 illustrates key trends, with the left panel presenting a "spaghetti plot" that shows CD4 counts over time for all individuals, while the right panel highlights selected trajectories, revealing both increasing and decreasing trends. Additionally, the average population trend of CD4 counts demonstrates an overall increase over time, although this approach does not capture dynamic changes at specific time points. Furthermore, a population-level view of CD4 counts relative to viral load suggests an inverse relationship; however, individual-level trends exhibit variability, with some showing positive or null associations. A linear regression across all points (not shown here for brevity) confirms this inverse relationship between CD4 counts and viral load, a finding of clinical significance for monitoring and managing patient responses.

It is important to analyze clinical responses to treatment over time, which can be achieved parametrically (e.g., using segmented or interrupted linear regression). This approach provides discrete coefficient estimates for each time interval, revealing that the CD4-viral load relationship is not constant across time points. However, this method does not produce a continuous trajectory for the relationship, resulting in a set of discrete coefficients for specific time points.

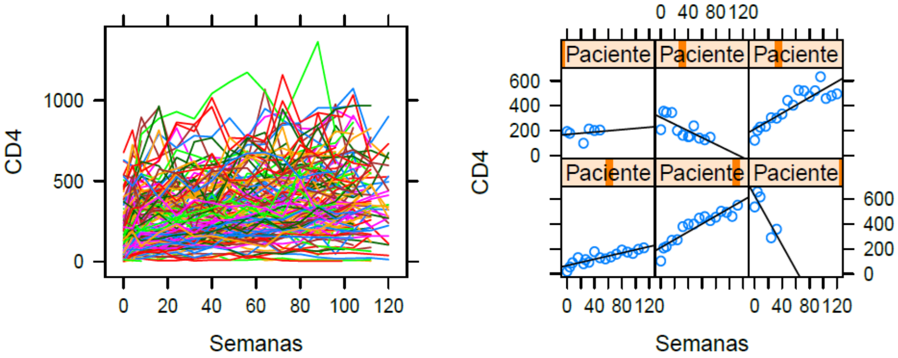


Figure 4: Spaghetti plot of the population and the linear trend of selected individuals' CD4 lymphocyte counts over time using data from ACTG-388.

To address this limitation, we propose a generalized non-parametric model to estimate a continuous curve of dynamic coefficients:

$$\log(\mathsf{E}(\mathrm{CD4}_{i,j})) = \beta_0(t_{i,j}) + \mathrm{RNA}(t_{i,j})\beta_1(t_{i,j}),\ j = 1, \ldots, n_i,\ i = 1, \ldots, n,$$

where $\mathrm{CD4}_{i,j}$ is the CD4 count, initially assumed to follow a Poisson distribution; $\mathrm{RNA}(t_{i,j})$ is the viral load (log scale) for the j-th measurement of the i-th patient; $\beta_0(t_{i,j})$ is the dynamic intercept coefficient, and $\beta_1(t_{i,j})$ is the dynamic coefficient for viral load. These coefficients can be estimated using either Kernel or Spline methods.

For the ACTG-315 dataset, we used the Akaike Information Criterion (AIC) to optimize smoothing parameters. The best Spline model utilized three knots per coefficient with a second-degree polynomial, while the Gaussian Kernel method employed bandwidths of 0.318 for β_0 and 0.241 for β_1, using five knots per coefficient. Model quality and fit were assessed using the Root Mean Square Error (RMSE) and AIC. Negative binomial distributions outperformed Poisson distributions in modeling the response variable for both Kernel and Spline approaches, as confirmed by the Voung test. However, the Voung test did not distinguish between the Spline and Kernel methods when using the negative binomial distribution. Residual deviance goodness-of-fit tests shown in Table 2 indicate no significant issues, suggesting that both Kernel and Spline models adequately fit the data without dispersion problems.

Figure 5 presents the estimated dynamic coefficients for $\beta(t)$ using Kernel and Spline methods. The dynamic slope, $\beta_1(t)$, represents the evolving relationship between CD4 counts and viral load. Bootstrap confidence intervals for $\beta(t)$ are shown in Figure 6. Both methods revealed a positive slope from the beginning until day 39, followed by a negative slope until day 118, and a return to a positive slope thereafter. This indicates treatment efficacy from day 39 to 118, during which CD4 counts increased, followed by reduced efficacy as CD4 counts declined. Kernel methods showed stabilization around day 157, while Spline methods oversmoothed

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

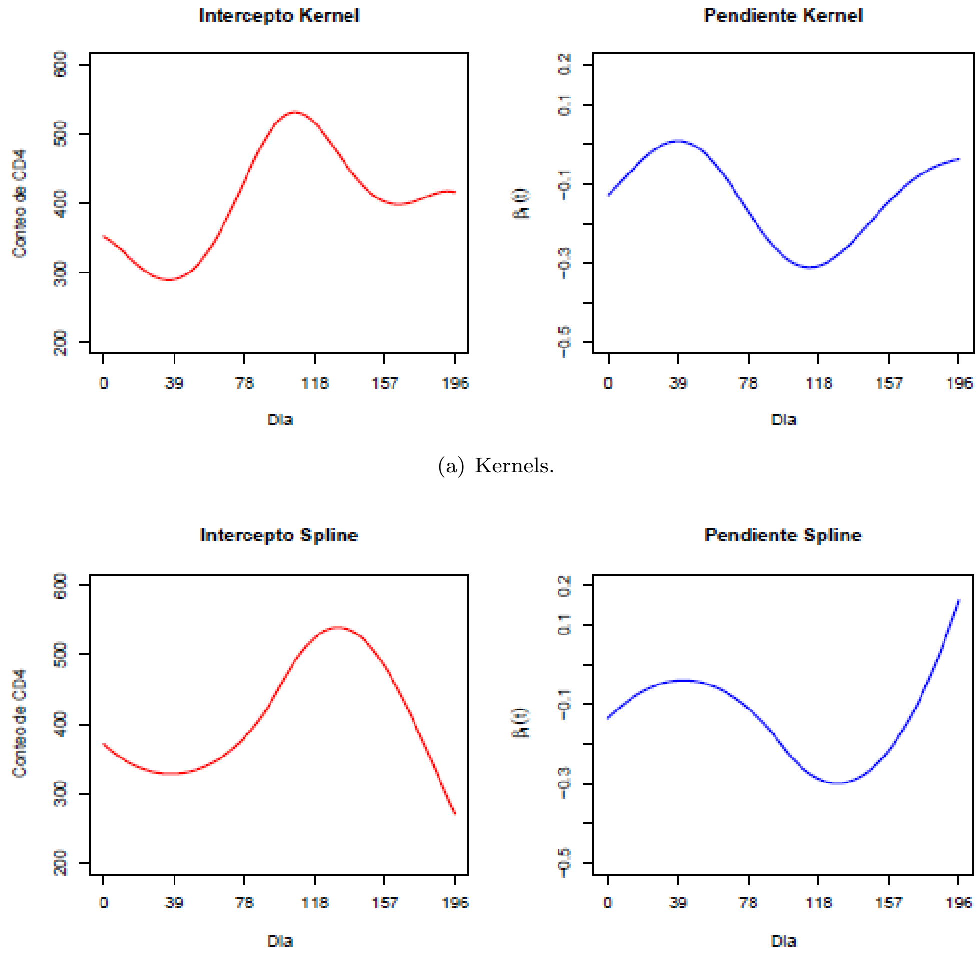
Table 2: Performance of the generalized model for estimating β with the ACTG-315 data.

the dynamics. In both methods, precision decreased near the study's end due to fewer observations, resulting in increased uncertainty.

In earlier studies, (Sosa and Díaz, 2009) modeled the relationship using regression splines and second-degree B-splines, finding weaker relationships early in treatment and stronger inverse relationships between days 100 and 140. Between days 140 and 150, the relationship appeared direct but weak. (Sosa and Buitrago, 2022) later applied Kernel and Spline regression to the ACTG-315 dataset (limited to a follow-up period of 86 days), identifying a direct relationship up to day 34, followed by a negative relationship until the end of the study period. Using the Deviance Information Criterion (DIC), they concluded that Kernel-based models offered the best performance. These findings emphasize the dynamic nature of the CD4-viral load relationship and the need for detailed temporal analyses.

In both clinical studies (ACTG-388 and ACTG-315), missing data primarily arose from dropouts or irregular attendance during follow-up. No imputation techniques were applied; instead, the estimation was carried out using only the available observations at each time point. The nonparametric models used in this study are inherently tolerant of incomplete data, as the estimation of the dynamic coefficients relies solely on observed values. This approach implicitly assumes that the missingness mechanism is at least missing at random (MAR). While this assumption is common in longitudinal analyses, future work should consider sensitivity analyses under different missingness scenarios and explore the use of multiple imputation methods when necessary.

In the HIV/AIDS applications, the dynamic patterns observed between viral load and CD4 lymphocyte counts provide clinically meaningful insights. For instance, identifying time intervals where the association between these markers is strongest can inform the optimal timing for clinical evaluations or treatment modifications. The ability to detect periods of declining immune response, even in the presence of antiretroviral therapy, may alert clinicians to the need for closer monitoring or alternative treatment strategies. These findings underscore the potential of time-varying coefficient models to support personalized patient management by adapting follow-up intensity according to the dynamic behavior of key clinical indicators.



(a) Kernels.

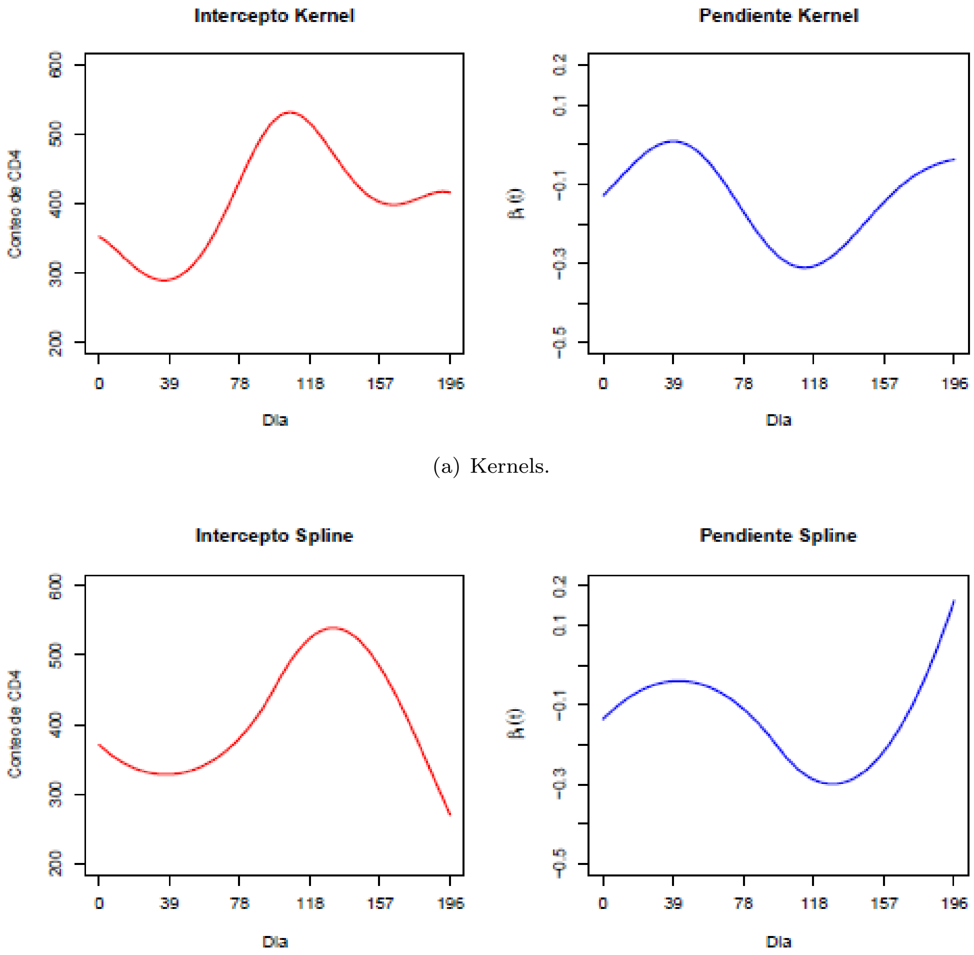
(b) Splines.

Figure 5: Estimated intercept β_0 and slope β_1 of CD4 lymphocyte counts using ACTG-315 data, with kernels and splines.

5.2. COVID-19 Case Study

In contrast to the analysis of the interplay between CD4 lymphocyte counts and viral load in HIV/AIDS, which focuses on understanding both population-level and individual dynamics over the study's time horizon, the analysis in the context of COVID-19 prioritizes predictive accuracy due to its critical role in informing public health policy decisions. Specifically, forecasting future case counts is essential for effective resource allocation and intervention planning.

The model applied to the COVID-19 case study follows the structure of model 4.17,



(a) Kernels.

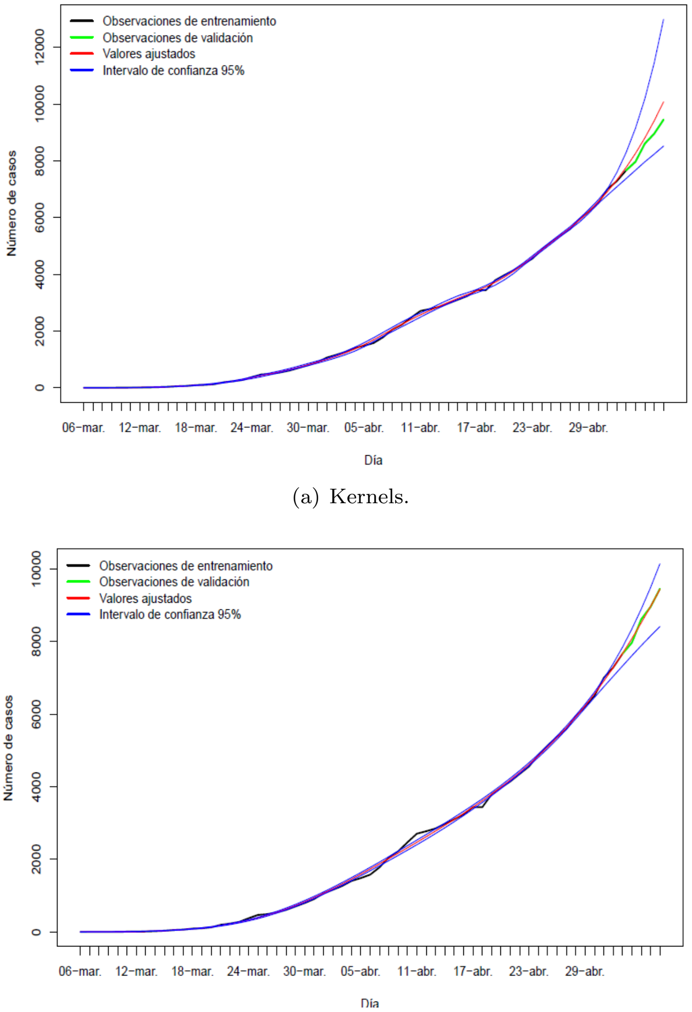
(b) Splines.

Figure 6: Confidence bands for the intercept β_0 and the slope β_1 of CD4 lymphocyte counts using ACTG-315 data, with kernels and splines.

incorporating a dynamic coefficient $\beta_0(t_{i,j})$ as the intercept without additional covariates. The logarithm of case counts is used as the response variable, which aligns with standard practices for modeling exponential growth patterns observed in infectious disease data.

To demonstrate the proposed estimation strategies and assess their predictive reliability, a training dataset consisting of observations from March 6, 2020, to May 2, 2020 (58 days), was used. A subsequent 5-day period was reserved for prediction and validation. The results of this analysis are illustrated in Figure 7, showing both model fit and predictive performance. Table 3 presents the evaluation metrics for models using different response variable formulations. Smoothing parameters were optimized using predictive cross-validation (PCV). The Spline-based approach employed a second-degree polynomial with 10 knots, while the Kernel-based method used 5 knots and a bandwidth of 0.7982983.

The Voung test was applied to compare the performance of these methodologies, revealing that Spline models outperformed Kernel models across both Poisson and negative binomial response functions. Furthermore, no statistically significant



(a) Kernels.

(b) Splines.

Figure 7: Confidence bands for the intercept β_0 of (log) number of active case using COVID-19 data, with kernels and splines.

Model	**AIC**	**RMSE**	$p(\chi^2)$
Kernel Poisson	624.1957	64.05078	1.853374×10^{-10}
Kernel NB	599.2826	77.23511	0.1790205
Spline Poisson	556.6131	44.03699	0.05398627
Spline NB	558.6182	44.03909	0.05455642

Table 3: Performance of the generalized dynamic coefficient models for estimating β with the COVID-19 data in Colombia from March to May 2020.

differences were found between the various Spline-based models. These findings highlight the robustness and accuracy of Spline methods in capturing the dynamics of COVID-19 case counts, making them a valuable tool for public health forecasting and decision-making.

In this case study, the nonparametric model captures the evolving trend in active cases with high resolution over short time horizons. This temporal sensitivity is particularly valuable for informing real-time public health decisions. For example, identifying turning points or changes in the growth rate of infections can guide the timing of interventions such as mobility restrictions, testing campaigns, or hospital resource planning. Although long-term forecasting remains limited due to the model's nonparametric nature, the ability to detect and interpret short-term dynamics makes this approach a useful tool for epidemic monitoring and adaptive policy response.

Finally, the use of predictive models in public health, particularly during a pandemic such as COVID-19, carries important ethical and operational implications. Forecasts derived from dynamic coefficient models can inform decisions about the timing and scale of interventions, such as lockdowns, vaccination strategies, or the deployment of medical resources. However, such predictions must be interpreted with caution and transparency, especially when they influence high-stakes decisions. Ensuring that model outputs are accompanied by appropriate uncertainty quantification and are communicated clearly to policymakers is essential to avoid misinterpretation or misuse. Additionally, using these tools to optimize resource allocation should be guided by principles of equity and public benefit, particularly in settings with limited healthcare capacity.

6. Discussion

This study develops generalized dynamic coefficient models for longitudinal data with applications in health sciences. Two smoothing strategies are employed: one using truncated power bases for splines and another using a Gaussian kernel. These approaches are proposed as alternatives to conventional parametric models, which estimate a finite and constant number of coefficients that may not adequately capture the dynamic nature of the response variable. Additionally, they serve as alternatives to non-parametric models that assume the response variable follows a normal distribution, which can be problematic when residuals exhibit biased behavior.

Building on issues identified in the statistical literature, particularly in prior studies (Huang et al., 2012; Lu and Huang, 2017), that modeled clinical trial data analyzed here using dynamic coefficients and assuming a normal distribution for the response variable (Sosa and Díaz, 2009; Huang and Lu, 2016), we propose alternative solutions to address potential biases in residual behavior. On one hand, the response variable was assumed to represent counts, and a Poisson distribution was applied to model its expected value. On the other hand, a generalized linear model was used to estimate dynamic coefficients with a known link function applied to the response variable.

Non-parametric dynamic coefficient models are rarely used in health research. In fact, during the review conducted for COVID-19—a highly topical issue—most models applied were parametric. The proposed models were demonstrated with data from the health sector. In one example, the dynamic relationship between two variables in HIV/AIDS was modeled, providing insight into internal data dynamics across the analyzed range, which is valuable for monitoring treatment. In another example, COVID-19 case predictions were made, illustrating the performance of the proposed models outside the data range by estimating extrapolable trends.

Regarding model performance, results from the ACTG-388 and ACTG-315 protocols indicated no significant differences between the kernel and spline methodologies when both employed a negative binomial link function. However, with COVID-19 data, the spline methodology outperformed the kernel approach. Still, no distinction was observed between Poisson and negative binomial link functions when smoothing with splines.

For the ACTG-388 data, CD4 lymphocyte counts showed a more pronounced increase during the first 40 weeks of treatment, with continued but slower increases until weeks 101–112, depending on the method used, after which the count declined. This suggests an enhanced treatment effect until nearly the end of the follow-up period, with a decline in the final weeks. In the ACTG-315 data, treatment effectiveness was observed from day 39 to day 118, beyond which effectiveness diminished.

The ACTG protocol data analyzed in this study has been extensively studied using other dynamic coefficient models, enabling comparisons with previous studies. Similar results were obtained, validating our models to some extent. However, a limitation of this study is the lack of access to national data, which would enable more immediate applications in local settings. Expanding the analysis to include more covariates would also be of interest.

For the COVID-19 case, due to the study's scope, only Colombian data were modeled, but extending the analysis to other countries and regions with repeated measures across multiple individuals would be highly relevant. Longer prediction periods could also be explored to determine the maximum time these models remain reliable and assess their ability to detect epidemic peaks. Comparing the proposed models with conventional parametric models, which are more commonly used to describe the pandemic, would further strengthen the analysis.

One limitation of TVCM models is that they estimate only the population's average behavior. Given that longitudinal data (LDA) includes repeated measures for each individual, extending the model to include mixed dynamic and random coefficients would allow trajectory estimation for clinical monitoring at the patient level. Another limitation is the lack of modeling for the correlation structure of repeated measures, which is essential as repeated observations within the same individual may be correlated. Non-parametric strategies do not require model specification and are particularly useful for capturing the non-linear dependence of the response variable on time-dependent covariates. However, they are less effective when covariates exhibit high dimensionality (Fan and Zhang, 2008).

A known limitation of nonparametric time-varying coefficient models is their limited capacity for extrapolation beyond the range of the observed data. Since these models rely on local smoothing techniques, their predictions tend to be reliable only within the support of the available observations. This is particularly relevant in forecasting contexts, such as epidemic modeling, where long-term projections are often needed. In such cases, caution must be exercised when interpreting trends near the boundaries of the data, as the models may not accurately capture future dynamics. To address this limitation, future work could explore hybrid strategies that combine nonparametric estimation with parametric components for improved extrapolation, or consider semiparametric approaches that impose structure in the tails while preserving flexibility within the data range.

From a methodological standpoint, nonparametric time-varying coefficient models differ substantially from conventional parametric approaches such as generalized linear models (GLMs), generalized estimating equations (GEEs), or mixed effects models. Parametric models assume a fixed functional form for the relationship between covariates and the response, which offers interpretability and computational efficiency but may lack flexibility when the true effect structure varies over time. In contrast, nonparametric models do not require prior specification of the functional form and instead estimate smooth trajectories directly from the data. This allows for greater adaptability in capturing nonlinear or time-modulated effects. However, nonparametric methods often require careful tuning of smoothing parameters and may be sensitive to sparsity in data across time. Additionally, while parametric models typically include explicit correlation structures for repeated measurements, the current nonparametric approach relies on independence assumptions, which may limit its efficiency in certain settings. These trade-offs underscore the importance of choosing modeling strategies based on the data characteristics and inferential objectives of each application.

Future extensions of this work could focus on several methodological and applied directions. Although the proposed nonparametric models provide flexible estimation of time-varying effects, they do not explicitly account for the correlation structure present in repeated measurements from the same individual. Incorporating random effects or modeling the intra-individual correlation structure, as done in mixed effects models or generalized estimating equations, could improve the efficiency of estimation and enable subject-specific inference. At present, the models are limited to capturing average population-level behavior. Extending the framework to include dynamic mixed effects models would allow the estimation of individual trajectories, which is particularly relevant in clinical contexts where patient-level monitoring is essential.

In addition, while the methods developed in this study perform well with a small number of covariates, their effectiveness may decrease in high-dimensional settings. This limitation suggests the need to explore variable selection procedures, regularization techniques, or dimension reduction approaches to achieve stable and efficient estimation. The current formulation does not include automatic variable selection or time-dependent interactions between predictors. In scenarios involving multiple covariates, it would be valuable to incorporate penalized estimation strategies and time-varying interaction terms to better capture complex temporal patterns. Finally, further developments could include alternative smoothing techniques and fully Bayesian formulations to enhance flexibility and improve uncertainty quantification. These extensions would broaden the applicability of the proposed models, for example by estimating antimicrobial resistance trajectories in pathogens, or by comparing disease dynamics across countries and regions in support of public health surveillance systems.

An additional limitation of this study is the absence of a direct comparison between the proposed models and conventional approaches widely used in longitudinal data analysis, such as generalized estimating equations (GEE) and generalized additive mixed models (GAMMs). While the primary goal was to illustrate the flexibility of nonparametric time-varying coefficient models in capturing nonlinear temporal relationships in count data, a comparative evaluation would provide a clearer understanding of their strengths and limitations relative to established parametric methods. As a direction for future research, we propose conducting a systematic comparison of predictive performance, individual-level fit, and the ability to model nonlinear effects across different modeling frameworks, which would contribute to a more robust empirical validation of the proposed approach.

Statements and Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this part.

Recibido: Diciembre de 2024
Aceptado: Julio de 2025

Referencias

I. Barbosa, K. de Lima, and A. de Almeida Medeiros. Covid-19 in brazil: Analysis of the pandemic short-term scenario in relation to other countries. *Int. J. Dev. Res*, 10(6):36840–36845, 2020.

M. Batista. Estimation of the final size of the covid-19 epidemic. *medRxiv*, 2020. doi: 10.1101/2020.02.16.20023606.

Z. Cai, J. Fan, and R. Li. Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, 95(451):888–902, 2000. doi: 10.1080/01621459.2000.10474248.

A. C. Cameron and P. K. Trivedi. *Regression Analysis of Count Data*, volume 53 of *Econometric Society Monographs*. Cambridge University Press, Cambridge, UK, 2nd edition, 2013. ISBN 978-0-521-18431-6.

J. Clara-Rahola. An empirical model for the spread and reduction of the covid-19 pandemic. *Studies of Applied Economics*, 38(2), 2020.

A. Cuevas. El análisis estadístico de grandes masas de datos: algunas tendencias recientes. *De la aritmética al análisis: historia y desarrollos recientes en matemáticas*, page 59, 2004.

M. Davidian, G. M. Fitzmaurice, G. Verbeke, and G. Molenberghs, editors. *Longitudinal Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL, 2008. ISBN 9781584886587.

D. de Pereda Sebastián, A. M. Ramos, and B. Ivorr. *Modelización matemática de la difusión de una epidemia de peste porcina entre granjas*. PhD thesis, Universidad Complutense de Madrid, 2010.

V. Díaz-Narváez, D. San-Martín-Roldán, A. Calzadilla-Núñez, P. San-Martín-Roldán, A. Parody-Muñoz, and G. Robledo-Veloso. Which curve provides the best explanation of the growth in confirmed covid-19 cases in chile? *Revista Latino-Americana de Enfermagem*, 28, 2020.

M. Egger, M. May, G. Chêne, A. N. Phillips, B. Ledergerber, F. Dabis, D. Costagliola, A. d. Monforte, F. De Wolf, P. Reiss, et al. Prognosis of hiv-1-infected patients starting highly active antiretroviral therapy: A collaborative analysis of prospective studies. *The Lancet*, 360(9327):119–129, 2002.

M. Ekum and A. Ogunsanya. Application of hierarchical polynomial regression models to predict transmission of covid-19 at global level. *Int. J. Clin. Biostat. Biom*, 6:027, 2020.

J. Fan and W. Zhang. Statistical methods with varying coefficient models. *Statistics and Its Interface*, 1(1):179–195, 2008. doi: 10.4310/SII.2008.v1.n1.a15.

J. J. Faraway. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* CRC Press, Boca Raton, FL, 2016. ISBN 978-1-4398-7442-0.

M. A. Fischl, M. Giuliano, S. Vella, H. J. Ribaudo, A. C. Collier, A. Erice, M. Dehlinger, J. J. Eron, M. S. Saag, S. M. Hammer, and G. D. Morse. 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, 2003. doi: 10.1086/377535.

N. Ford, K. Stinson, H. Gale, E. J. Mills, W. Stevens, M. P. González, J. Markby, and A. Hill. Cd4 changes among virologically suppressed patients on antiretroviral therapy: A systematic review and meta-analysis. *Journal of the International AIDS Society*, 18(1):20061, 2015.

E. Gonzalo-Gil, U. Ikediobi, and R. E. Sutton. Mechanisms of virologic control and clinical characteristics of hiv+ elite/viremic controllers. *The Yale Journal of Biology and Medicine*, 90(2):245–259, 2017.

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993.

D. R. Hoover, J. A. Rice, C. O. Wu, and L.-P. Yang. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822, 1998. doi: 10.1093/biomet/85.4.809.

Y. Huang and T. Lu. Bayesian inference on partially linear mixed-effects joint models for longitudinal data with multiple features. *Computational Statistics*, 32, 2016. doi: 10.1007/s00180-016-0682-9.

Y. Huang, J. Chen, and C. Yan. 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), 2012. doi: 10.1515/1557-4679.1364.

M. D. Hughes, V. A. Johnson, M. S. Hirsch, J. W. Bremer, T. Elbeik, A. Erice, D. R. Kuritzkes, W. A. Scott, S. A. Spector, N. Basgoz, et al. 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, 1997.

D. Kasilingam, S. P. Sathiya Prabhakaran, D. K. Rajendran, V. Rajagopal, T. Santhosh Kumar, and A. Soundararaj. 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*, 2020.

E. L. Korenromp, B. G. Williams, G. P. Schmid, and C. Dye. Clinical prognostic value of rna viral load and cd4 cell counts during untreated hiv-1 infection: A quantitative review. *PloS One*, 4(6):e5950, 2009.

N. Kumar. Review of innovation diffusion models. 2015.

M. M. Lederman, E. Connick, A. Landay, D. R. Kuritzkes, J. Spritzler, M. St. Clair, B. L. Kotzin, L. Fox, M. Heath Chiozzi, J. M. Leonard, et al. 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, 1998. doi: 10.1086/515608.

Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong, et al. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 382(13): 1199–1207, 2020.

H. Liang, H. Wu, and R. J. Carroll. 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, 2003.

S. R. Lipsitz, J. Ibrahim, and G. Molenberghs. 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, 2000. doi: 10.1111/1467-9876.00234.

T. Lu and Y. Huang. 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, 2017.

Y. Lu and R. Zhang. Smoothing spline estimation of generalized varying-coefficient mixed model. *Journal of Nonparametric Statistics*, 21(7):815–825, 2009. doi: 10.1080/10485250903020502.

J. Luo. Predictive monitoring of covid-19. *SUTD Data-Driven Innovation Lab*, 2020.

J. Ma. Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling*, 5:129–141, 2020.

F. Manrique Abril, V. M. González-Chordá, O. A. Gutiérrez Lesmes, C. F. Tellez Piñerez, and G. M. Herrera-Amaya. Modelo sir de la pandemia de covid-19 en colombia. 2020.

I. C. Marschner, A. C. Collier, R. W. Coombs, R. T. D'Aquila, V. DeGruttola, M. A. Fischl, S. M. Hammer, M. D. Hughes, V. A. Johnson, D. A. Katzenstein, et al. 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, 1998.

M. Mercker, U. Betzin, and D. Wilken. What influences covid-19 infection rates: A statistical approach to identify promising factors applied to infection data from germany. 2020.

J. Millán-Oñate, A. J. Rodríguez-Morales, G. Camacho-Moreno, H. Mendoza-Ramírez, I. A. Rodríguez-Sabogal, and C. Álvarez-Moreno. A new emerging zoonotic virus of concern: the 2019 novel coronavirus (sars-cov-2). *Infectio*, 24 (3):187–192, 2020.

Ministerio de Salud y Protección Social. *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.* Ministerio de Salud y Protección Social, Bogotá, Colombia, 2014.

D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introducción al análisis de regresión lineal.* Limusa Wiley, México, 2006. ISBN 978-9681859246.

W. H. Organization. *WHO case definitions of HIV for surveillance and revised clinical staging and immunological classification of HIV-related disease in adults and children.* World Health Organization, Geneva, Switzerland, 2007.

T. Park and S. Jeong. Analysis of poisson varying-coefficient models with autoregression. *Statistics*, 52(1):34–49, 2018. doi: 10.1080/02331888.2017.1405822.

E. Pelinovsky, A. Kurkin, O. Kurkina, M. Kokoulina, and A. Epifanova. Logistic equation and covid-19. *Chaos, Solitons & Fractals*, 140:110241, 2020.

J. Ramsay, G. Hooker, and S. Graves. *Functional Data Analysis with R and MATLAB.* Springer Science & Business Media, New York, NY, 2009. ISBN 978-0-387-98184-0.

J. Sosa and L. Buitrago. Time-varying coefficient model estimation through radial basis functions. *Journal of Applied Statistics*, 49(10):2510–2534, 2022.

J. Sosa and L. Díaz. *Desarrollo de un modelo de coeficientes dinámicos y aleatorios para el análisis longitudinales.* Master's thesis, Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia, 2009.

J. C. Sosa and L. G. Díaz. Random time-varying coefficient model estimation through radial basis functions. *Revista Colombiana de Estadística*, 35(1):167–184, 2012.

R. Thiébaut, P. Morlat, H. Jacqmin-Gadda, D. Neau, P. Mercié, F. Dabis, G. Chêne, et al. Clinical progression of hiv-1 infection according to the viral response during the first year of antiretroviral treatment. *AIDS*, 14(8):971–978, 2000.

C. H. S. Trujillo. 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), 2020.

J. W. Twisk. *Applied longitudinal data analysis for epidemiology: a practical guide.* Cambridge University Press, 2013.

Y. Wang. *Varying-Coefficient Models: New Models, Inference Procedures, and Applications*. Cambridge University Press, Cambridge, UK, 2007. ISBN 978-0-521-85936-9.

C. O. Wu and X. Tian. *Nonparametric Models for Longitudinal Data: With Implementation in R*. Chapman and Hall/CRC, Boca Raton, FL, 2018. ISBN 978-1-4987-5123-4.

H. Wu and J.-T. Zhang. *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches*, volume 515. John Wiley & Sons, Hoboken, NJ, 2006. ISBN 978-0-471-69982-0.

S. L. Zeger and P. J. Diggle. Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, 50(3):689–699, 1994. doi: 10.2307/2532807.

D. Şentürk and H.-G. Müller. Generalized varying coefficient models for longitudinal data. *Biometrika*, 95(3):653–666, 2008. doi: 10.1093/biomet/asn034.