UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura

Métodos de estimación para el modelo lineal funcional truncado

Federico Tomás Choque

Directora: Dra. Daniela Rodríguez

27 de Marzo de 2023

Agradecimientos

Todos estos años de estudio no hubiesen llegado a su fin de no ser por mi familia, que a pesar de que ser algo distante con ellxs, fueron mi motor con su apoyo, amor y paciencia en los momentos más complicados. Mis más sinceros agradecimientos a mi mamá y papá, cuya tenacidad para perseguir sus sueños yo intento replicar en los míos. A Anto, Fiore y Caro porque a pesar de ser el único varón, siempre me tuvieron en cuenta en las reuniones y eventos que hacían. A Narella, por sumar mucha alegría en estos últimos meses. A mis familiares peluditos y también a aquellxs que ya no están.

A lxs pibxs del Remando en el CBC, por permitirme participar de este lindo proyecto. Gracias Lucila, Ceci, Lau, Rodri, Mathi, David, Lenny, Pedro, Nahuel y Alexis por regalarme momentos para compartir debates y chismes, por ser varixs también compañerxs de cursada y por los viernes de catarsis que tanto necesitamos en su momento.

Gracias a mis compas de varias materias por formar parte de mi formación y dejarme también ser parte de la de ustedes: Alina Chocrón, Nico Marucho, Cami Capdevila, Eze Salvatierra y Aye Campero. No puedo olvidarme de los del CBC, los primeros con los que hablé y que hicieron que durante el primer año no me sienta solo: Juli Braier y Alex Chandia.

A Maxi Silva, a quien ya conocía de antes pero tuve el placer de conocer mejor en la facultad. Gracias por la amabilidad y simpatía en todos estos años.

A los de siempre: Facu y Lauti por mantener la amistad intacta durante todo este tiempo posterior al secundario.

A la Universidad de Buenos Aires, al Instituto Bunge, a la Universidad de San Andrés y a la Universidad de General Sarmiento por permitirme trabajar allí, en especial a los primeros dos por también formarme en mis estudios y volverme a abrir las puertas.

A lxs docentes del Instituto Bunge y a los del CBC de la sede Moreno por la calidez y el buen trato recibido durante mi trabajo con ellxs.

Agradezco a lxs docentes que tuve durante mi paso por Exactas, en especial a Ana Bianco, Gonzalo Chebi y Daniela Rodríguez por introducirme a esta área de estudio con mucha claridad, entusiasmo y amabilidad. Por responder mis preguntas y curiosidades que me surgían respecto cómo seguir mi plan de estudio.

Gracias Ana y Marina por formar parte del jurado y a Dani por ser parte de este cierre dirigiendo la presente tesis y del comienzo de otra etapa permitiéndome ingresar a la Maestría en Estadística Matemática del Instituto de Cálculo.

Índice general

Capítulo 1

Introducción

A la Estadística se le asocia el estudio de vectores observados $x_1, x_2, \ldots, x_n \in \mathbb{R}^p$ sobre los cuales se intenta inferir el mecanismo de la variable/vector aleatorio que los genera. Estos vectores representan, por ejemplo, los individuos en un estudio médico o unidades experimentales. Los valores ordenados de cada vector/individuo, en cambio, representan las mediciones de diferentes variables, como pueden ser el nivel de glucosa en sangre, la altura de una persona, la temperatura, etc. Durante el siglo XX, la teoría sobre inferencia estadística se desarrolló asumiendo $p < n$ y muchos resultados se obtuvieron basándose en que la cantidad de variables (p) queda fija mientras que la cantidad de experimentos (n) se hace infinitamente grande. Sin embargo, el análisis en los casos en los que p es relativamente 'grande' es cada vez más habitual y trae consigo el problema de la maldición de la dimensión, por lo cual requiere un enfoque distinto al tradicional.

En algunos casos es posible tratarlo mediante el análisis de datos funcionales. Desde esta perspectiva, los datos observados serán funciones y provendrán de una variable aleatoria funcional. Por ejemplo, la altura de una persona a lo largo de su vida que puede registrarse diariamente, siendo la cantidad de registros diarios mucho mayor que la del número de individuos por lo que se puede pensar que los datos provienen de una función respecto del tiempo $f : [0, T] \to \mathbb{R}_{\geq 0}$. Otro problema es el de escrituras de texto a mano: para cada registro, en lugar de identificar cada píxel del cuadro de texto como una variable distinta con valor en el intervalo $[0, 1]$ (según la intensidad del trazo), se puede pensar a cada muestra como proveniente de una función $f : [a, b] \times [c, d] \to [0, 1]$.

Muchas de las técnicas de aprendizaje supervisado como no supervisado conocidas en inferencia finito-dimensional tienen su versión funcional como el análisis de componentes principales, correlación canónica, análisis discriminante, entre otros. Esta tesis se enfocará en los modelos de regresión lineal para el caso funcional. En particular, se analizará el llamado *modelo lineal truncado para datos funcionales* que se caracteriza por incorporar un instante desconocido en el cual las observaciones dejan de ser relevantes para predecir las respuestas.

El término *datos funcionales* recién surgía por los 80'. En varios artículos ya se mencionaba el nuevo paradigma respecto a cómo afrontar los problemas de los datos de alta dimensión como Ramsay (1982) [18] y Dalzell y Ramsay (1991) [5] y también

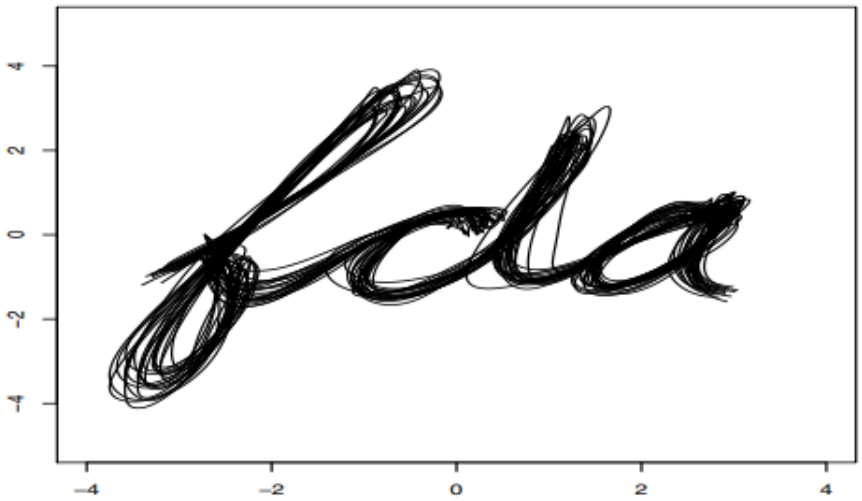


Figura 1.1: 20 muestras de escritura de la palabra *fda*. La unidad de los ejes está en centímetros.

sus aplicaciones como en Gasser et al. (1990) [8] en su análisis para el crecimiento de niños. A pesar de su auge en esos años, durante aquel siglo aparecieron diversos resultados como en Karhunen (1946) [13], Grenander (1950) [10] y Kleffe (1973) [14] sobre formalidades en la inferencia en espacios de Hilbert abstractos y componentes principales para datos funcionales. A fines de siglo XX, ya se comenzaba a analizar modelos de regresión lineal aplicados a datos funcionales como en Ramsay et al. (1997, first edition) [17] y Cardot et al. (1999) [3]. Malfait et al. (2003) [16] trataron el modelo lineal funcional histórico en el que las respuestas son funcionales y se intenta determinar el instante donde la función de peso bivariada se anula. Hall et al. (2015) analizan el caso con respuesta escalar mediante dos métodos. Otras restricciones de forma sobre la función de peso para respuestas escalares, como la monotonía o la convexidad, son analizadas en el trabajo de Benjamín (2020) [1].

El objetivo del presente trabajo es analizar los métodos propuestos por Hall y Hooker, comparar su efectividad a través de los diferentes parámetros y las desventajas surgidas ante diferentes modelos. En el capítulo 2 se presentan las definiciones, medidas y resultados elementales del Análisis de Datos Funcionales (FDA de ahora en adelante) a partir del Análisis Funcional. En el capítulo 3 se presentan los modelos lineales en el contexto de los datos funcionales, su comparación con el caso finito-dimensional y diferentes formas de estimar sus parámetros. En el capítulo 4 se introduce el modelo lineal funcional truncado, su buena definición y los dos métodos propuestos por Hall y Hooker. En el capítulo 5 figuran los resultados de las simulaciones para diferentes modelos y un análisis de lo obtenido.

Capítulo 2

Datos funcionales

El objetivo de este capítulo es presentar las definiciones, herramientas y resultados que se usan en el FDA. Todo esto en el contexto del Modelo Lineal Funcional por lo que es posible que no figuren los resultados más importantes del FDA sino aquellos que ayuden a una mejor comprensión para la teoría del capítulo siguiente. Se brindan las definiciones de las unidades básicas de estudio como en Ferraty et al. (2006) [7].

Definición 2.0.1. *Dado un espacio de probabilidad $(\Omega, \mathcal{F}, P)$, y un elemento aleatorio $X : \Omega \to \mathcal{H}$, se dice que es funcional si $\mathcal{H}$ es un espacio de dimensión infinita.*

Notemos que si $\mathcal{H}$ es $\mathbb{R}$ o $\mathbb{R}^p$ entonces se trata de las tradicionales variables aleatorias y vectores aleatorios. $\mathcal{H}$ podría ser un espacio de curvas o de superficies dotado de un producto interno, en nuestro caso consideramos $\mathcal{H} = L^2(I)$ con I un intervalo real. Esto se puede generalizar para cualquier otro espacio de Hilbert de dimensión infinita.

Definición 2.0.2. *Se dice que $x_1, \ldots, x_n$ forman un conjunto de datos funcionales si son observaciones de n funcionales aleatorios $X_1, \ldots, X_n$ idénticamente distribuidos.*

2.1. Elementos sobre espacios de Hilbert

Previo a generalizar las medidas conocidas sobre datos en $\mathbb{R}$ o $\mathbb{R}^p$ a un espacio de Hilbert infinito dimensional, necesitamos revisar algunas definiciones y resultados sobre teoría de operadores. Se podrá observar que algunos resultados ya eran conocidos para el caso de matrices pero que en estos espacios se necesitan hipótesis adicionales

Definición 2.1.1. *Dado un espacio de Hilbert $\mathcal{H}$, decimos que un operador lineal $T : \mathcal{H} \to \mathcal{H}$ es acotado (o continuo) si $\{||T(x)|| : ||x|| \leq 1\}$ está acotado superiormente. En tal caso, definimos la norma del operador como*

$$||T|| := \sup_{||x|| \leq 1} ||T(x)||.$$

Definición 2.1.2. *Dado un espacio de Hilbert $\mathcal{H}$, decimos que un operador lineal $T : \mathcal{H} \to \mathcal{H}$ es compacto si $\overline{T(B)}$ es compacto para todo $B \subset \mathcal{H}$ acotado.*

Se deduce de la definición que un operador lineal compacto es también acotado.

Definición 2.1.3. *Dado un espacio de Hilbert $\mathcal{H}$, y un operador lineal $T : \mathcal{H} \to \mathcal{H}$ decimos que $S : \mathcal{H} \to \mathcal{H}$ es adjunto de T si para todo $x, y \in \mathcal{H}$ se tiene*

$$\langle T(x), y\rangle = \langle x, S(y)\rangle.$$

Se lo nota T^*.

Observación 2.1.1. *T^* también es lineal.*

Notemos que esto generaliza el operador trasposición de una matriz $A \in \mathbb{R}^{n\times m}$. Esto pues para el producto interno canónico y $T(x) = Ax$, se tiene que $T^*(x) = A^T x$. En general, el adjunto de un operador podría no existir pero se puede asegurar su existencia y unicidad bajo ciertas condiciones. Para eso primero enunciemos el teorema de Riesz que ayudará en la demostración de lo anterior.

Teorema 2.1.1. *(Teorema de representación de Riesz-Fréchet) Dado un espacio de Hilbert $\mathcal{H}$, y un operador lineal acotado $T : \mathcal{H} \to \mathbb{R}$ entonces existe un único $y \in \mathcal{H}$ tal que para todo $x \in \mathcal{H}$ se tiene $T(x) = \langle x, y\rangle$.*

Demostración. Ver teorema 10.2.3 de Kokoszka (2017) [15]

Proposición 2.1.2. *Dado un espacio de Hilbert $\mathcal{H}$, y un operador lineal acotado $T : \mathcal{H} \to \mathcal{H}$ entonces existe su adjunto y es único.*

Demostración. Para cada $v \in \mathcal{H}$, definimos el operador lineal $\varphi_v : \mathcal{H} \to \mathcal{H}$ como $\varphi_v(x) = \langle T(x), v\rangle$. Notemos que φ_v es acotado, que se deduce de la desigualdad de Cauchy-Schwarz y de que T es acotado. Por el teorema de Riesz, entonces existe un único $y_v \in \mathcal{H}$ (pues depende de v) tal que para todo $x \in \mathcal{H}$ se tiene

$$\langle T(x), v\rangle = \varphi_v(x) = \langle x, y_v\rangle.$$

Basta tomar entonces $T^*(v) = y_v$ que está bien definida pues y_v queda unívocamente determinado por el teorema de Riesz.

Para ver la unicidad, supongamos que existe otro adjunto $S : \mathcal{H} \to \mathcal{H}$. Luego se tiene que para todo $x, y \in \mathcal{H}$

$$\begin{aligned}\langle x, T^*(y)\rangle &= \langle x, S(y)\rangle\\ \langle x, T^*(y) - S(y)\rangle &= 0\\ \langle x, (T^* - S)(y)\rangle &= 0\end{aligned}$$

.

En consecuencia, $T^* = S$. □

Mientras que una matriz se decía que era simétrica si era igual a su traspuesta, para operadores se define lo siguiente.

Definición 2.1.4. *Dado un espacio de Hilbert $\mathcal{H}$, decimos que un operador lineal acotado $T : \mathcal{H} \to \mathcal{H}$ es autoadjunto si $T = T^*$.*

Antes de demostrar la generalización de la descomposición espectral de matrices simétricas, precisamos el siguiente resultado.

Teorema 2.1.3. *(Teorema de Hilbert-Schmidt) Dado un espacio de Hilbert $\mathcal{H}$, y un operador lineal compacto y autoadjunto $T : \mathcal{H} \to \mathcal{H}$ entonces existe un sistema ortonormal de autovectores de T $\{u_j\}_{j\in\mathbb{N}}$ de autovalores no nulos tales que todo $x \in \mathcal{H}$ admite una escritura de la siguiente forma*

$$x = \sum_{j=1}^{\infty} \alpha_j u_j + v,$$

donde $\alpha_j \in \mathbb{R}$ y $v \in \mathcal{H}$.

Demostración. Ver teorema 4.10.1 de Debnath (2005) [6].

Observemos que si una matriz cuadrada $A \in \mathbb{R}^{n\times n}$ es simétrica, entonces admite una base ortonormal de autovectores $\{v_j\}_{j=1}^n$ con $\{\lambda_j\}_{j=1}^n$, sus respectivos autovalores tal que A se escribe como

$$A = \begin{pmatrix} v_1 & v_2 & \cdots & v_n \end{pmatrix} \begin{pmatrix} \lambda_1 & & & \\ & \ddots & & \\ & & & \lambda_n \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{pmatrix} = \sum_{j=1}^{n} \lambda_j v_j v_j^T.$$

Escrito como una transformación lineal:

$$Ax = \sum_{j=1}^{n} \lambda_j x^T v_j \cdot v_j.$$

El siguiente teorema introduce un resultado similar a esta escritura pero para operadores.

Teorema 2.1.4. *(Teorema espectral para operadores compactos autoadjuntos) Dado un espacio de Hilbert $\mathcal{H}$, y un operador lineal compacto y autoadjunto $T : \mathcal{H} \to \mathcal{H}$ entonces existe un sistema ortonormal completo de autovectores $\{u_j\}_{j\in\mathbb{N}}$ con $\{\lambda_j\}_{j\in\mathbb{N}}$, sus respectivos autovalores tal que T se escribe como*

$$T(x) = \sum_{j=1}^{n} \lambda_j \langle x, u_j\rangle \cdot u_j.$$

Demostración. Ver corolario 4.10.2 de Debnath (2005) [6].

Definición 2.1.5. *Dado un espacio de Hilbert $\mathcal{H}$, y un operador lineal autoadjunto $T : \mathcal{H} \to \mathcal{H}$, decimos que es semidefinido positivo si $\langle T(v), v\rangle \geq 0$ para todo $v \in \mathcal{H}$.*

Si bien, la teoría vista vale en cualquier espacio de Hilbert infinito-dimensional, nos detendremos en $L^2(I)$ que es en donde se centra esta tesis. Definimos allí el siguiente tipo de operadores.

Definición 2.1.6. *Dado $\psi \in L^2(I \times I)$, definimos el operador integral $\Psi : L^2(I) \to L^2(I)$ como*

$$\Psi(f)(t) = \int_I \psi(t, s) f(s) ds.$$

En tal caso decimos que ψ es el núcleo del operador Ψ.

Proposición 2.1.5. *Para un operador integral como el anterior:*

- *Ψ es compacto.*
- *Si el núcleo ψ es simétrico, entonces Ψ también es autoadjunto y semidefinido positivo.*

Teorema 2.1.6. *(Teorema de Mercer) Dado $\psi \in L^2(I \times I)$ continua y simétrica, entonces*

$$\psi(s, t) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(s)\varphi_j(t),$$

donde $\{\varphi_j\}$ es una base ortonormal de autovectores del operador integral de núcleo ψ y $\{\lambda_j\}$, sus respectivos autovalores ordenados de mayor a menor. Más aún, la convergencia de la serie es uniforme.

2.2. Medidas de resumen

Las medidas conocidas para el caso finito-dimensional se redefinen en este contexto. Se las define a partir de las medidas de sus proyecciones por lo que todas involucran el producto interno.

2.2.1. Esperanza

Definición 2.2.1. *Dado un espacio de Hilbert $\mathcal{H}$ y X un funcional aleatorio en $\mathcal{H}$, decimos que es integrable si $\mathbb{E}(||X||) < \infty$.*

Definición 2.2.2. *Dado un espacio de Hilbert $\mathcal{H}$ y X un funcional aleatorio en $\mathcal{H}$ integrable, definimos su esperanza como aquel $\mu \in \mathcal{H}$ que para todo $v \in \mathcal{H}$ satisface*

$$\langle \mu, v \rangle = \mathbb{E}(\langle X, v \rangle).$$

Lo notamos $\mathbb{E}(X)$.

Observación 2.2.1. *La existencia y unicidad de la esperanza está bien definida. Se deduce de aplicar el teorema de Riesz sobre el operador lineal acotado $T : \mathcal{H} \to \mathbb{R}$ dado por $T(v) = \mathbb{E}(\langle X, v \rangle)$.*

Observemos cómo esta definición para el caso funcional es consistente con la versión finito-dimensional. Recordemos que la esperanza de un vector aleatorio $X = (X_1, \ldots, X_p)$ era otro vector $\mu \in \mathbb{R}^p$ donde $\mu_i = \mathbb{E}(X_i)$ para $i = 1, \ldots, p$. Usando el producto interno canónico y denotando e_i al iésimo vector canónico de $\mathbb{R}^p$, se obtiene la definición original a partir de la versión funcional con $v = e_i$ para $i = 1, \ldots, p$.

$$\begin{aligned} \langle \mu, e_i \rangle &= \mathbb{E}(\langle X, e_i \rangle), \\ \mu_i &= \mathbb{E}(X_i). \end{aligned}$$

Proposición 2.2.1. *Sea X un funcional aleatorio en $L^2(I)$ integrable, entonces se tiene que $\mathbb{E}(X)(t) = \mathbb{E}(X(t))$, donde la igualdad es en L^2.*

Demostración. Ver ejemplo 11.2.1 de Kokoszka (2017) [15].

Observemos que si definimos, de manera análoga al caso finito-dimensional, la media muestral como $\overline{X}_n := \frac{1}{n} \sum_{j=1}^{n} X_j$, éste resulta ser un estimador insesgado y (débilmente) consistente para la esperanza. Esto se obtiene a partir del siguiente resultado. Para eso, usemos el siguiente lema.

Definición 2.2.3. *Dado un espacio de Hilbert $\mathcal{H}$ y X un funcional aleatorio en $\mathcal{H}$, decimos que es de cuadrado integrable si $\mathbb{E}(||X||^2) < \infty$.*

Lema 2.2.2. *Sean X_1 y X_2 funcionales aleatorios en $L^2(I)$ de cuadrado integrable e independientes tal que $\mathbb{E}(X_1) = 0$, entonces se tiene que $\mathbb{E}(\langle X_1, X_2\rangle) = 0$.*

Teorema 2.2.3. *Sean $(X_j)_{j\in\mathbb{N}}$, una sucesión de funcionales aleatorios en $L^2(I)$ de cuadrado integrable, i.i.d. de esperanza μ, entonces $\mathbb{E}(\overline{X}_n) = \mu$ y $\mathbb{E}(||\overline{X}_n - \mu||^2) = O(\frac{1}{n})$.*

Demostración. Ver teorema 2.3 de Horváth (2012) [12].

2.2.2. Covarianza

Definición 2.2.4. *Dado un espacio de Hilbert $\mathcal{H}$ y X un funcional aleatorio en $\mathcal{H}$ de cuadrado integrable, definimos su operador de covarianza como operador lineal $C : \mathcal{H} \to \mathcal{H}$ dado por*

$$C(v) = \mathbb{E}\left(\langle X - \mu, v\rangle (X - \mu)\right).$$

Observación 2.2.2. *Para todo $u, v \in \mathcal{H}$, se tiene que*

$$\langle C(v), u\rangle = \mathbb{E}\left(\langle X - \mu, v\rangle, \langle X - \mu, u\rangle\right) = \mathrm{Cov}(\langle X, v\rangle, \langle X, u\rangle).$$

En el caso $\mathbb{R}^p$, la covarianza para un vector aleatorio $X = (X_1, \ldots, X_p)$, es una matriz $\mathbb{R}^{p\times p}$ definida como $(\mathrm{Cov}(X))_{ij} := \mathrm{Cov}(X_i, X_j)$. Como a cualquier matriz, se la puede considerar como una trasformación lineal $\mathrm{Cov}(X) : \mathbb{R}^p \to \mathbb{R}^p$. Usando el producto interno canónico se obtiene la definición original a partir de la observación anterior con $v = e_i$ y $u = e_j$ para $i, j = 1, \ldots, p$.

$$\begin{aligned}\langle \mathrm{Cov}(X)(e_i), e_j\rangle &= \mathrm{Cov}(\langle X, e_i\rangle, \langle X, e_j\rangle)\\ e_i^T \,\mathrm{Cov}(X) e_j &= \mathrm{Cov}(X_i, X_j)\\ (\mathrm{Cov}(X))_{ij} &= \mathrm{Cov}(X_i, X_j).\end{aligned}$$

Proposición 2.2.4. *Sea X un funcional aleatorio en $\mathcal{H}$ de cuadrado integrable, entonces*

- *C es autoadjunto,*
- *C es semidefinido positivo,*
- *los autovalores λ_j de C cumplen $\sum_j \lambda_j < \infty$ y*
- *C es compacto.*

Como consecuencia del teorema de descomposición espectral, se tiene el siguiente resultado.

Proposición 2.2.5. *Sea X un funcional aleatorio en $\mathcal{H}$ de cuadrado integrable, entonces existe una base ortonormal de autovectores $\{u_j\}_{j\in\mathbb{N}}$ con $\{\lambda_j\}_{j\in\mathbb{N}} \subset \mathbb{R}_{\geq 0}$, sus respectivos autovalores ordenados de mayor a menor tal que $\lambda_j \to 0$ y C se escribe como*

$$C(v) = \sum_{j=1}^{\infty} \lambda_j \langle v, u_j \rangle \cdot u_j.$$

Volvemos al contexto de $L^2(I)$ donde obtenemos una reescritura del operador de covarianza.

Proposición 2.2.6. *Sea X un funcional aleatorio en $L^2(I)$ de cuadrado integrable, entonces el operador de covarianza $C : L^2(I) \to L^2(I)$ se escribe como*

$$C(f)(t) = \int_I c(t,s) f(s) ds$$

donde $c(t,s) = \mathrm{Cov}(X(t), X(s))$.

Demostración. Suponiendo que X está definido en un espacio de probabilidad $(\Omega, \mathcal{F}, P)$, sean $f, g \in L^2(I)$, usando la observación 2.2.2 se tiene lo siguiente:

$$\begin{aligned}
\langle C(f), g \rangle &= \mathbb{E}\left(\langle X-\mu, f\rangle, \langle X-\mu, g\rangle\right) = \int_\Omega \langle X-\mu, f\rangle \langle X-\mu, g\rangle dP \\
&= \int_\Omega \left(\int_I (X(s)-\mu(s)) f(s) ds \right) \left(\int_I (X(t)-\mu(t)) g(t) dt \right) dP \\
&= \int_\Omega \left(\int_I \int_I (X(t)-\mu(t))(X(s)-\mu(s)) f(s) g(t) ds dt \right) dP \\
&= \int_I \int_I \left(\int_\Omega (X(t)-\mu(t))(X(s)-\mu(s)) dP \right) f(s) g(t) ds dt \\
&= \int_I \int_I c(t,s) f(s) g(t) ds dt = \left\langle \int_I c(\cdot, s) f(s) ds, g \right\rangle,
\end{aligned}$$

donde la quinta igualdad se desprende de aplicar el teorema de Fubini a la función $(\omega, t, s) \longmapsto (X_\omega(t) - \mu(t))(X_\omega(s) - \mu(s))$ que está en el espacio $L^1(\Omega \times I \times I)$. □

Sin embargo, se suele trabajar con el núcleo en lugar del operador de covarianza. Notemos que si es continua, por el teorema de Mercer el núcleo admite un desarrollo en serie de autofunciones ordenadas del operador de covarianza que forman una base ortonormal donde la convergencia es uniforme en I. Esas autofunciones son las llamadas componentes principales funcionales. El siguiente resultado que es una variante del teorema de Hilbert-Schmidt permite describir a un funcional aleatorio a partir de su media y autofunciones del operador de covarianza.

Proposición 2.2.7. *(Desarrollo de Karhunen-Loéve) Sea X un funcional aleatorio en $L^2(I)$ de cuadrado integrable con núcleo de covarianza continuo, entonces se tiene que*

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_j \varphi_j(t),$$

donde $\mu = \mathbb{E}(X)$ y $\{\varphi_j\}_j$ son las autofunciones del operador de covarianza ordenados de mayor a menor según el autovalor λ_j y $\xi_j = \langle X - \mu, \varphi_j \rangle$. Además

- *La serie converge uniformemente en $L^2(I)$ en el siguiente sentido:*

$$\lim_{p\to\infty} \sup_{t\in I} \mathbb{E}\left(X(t) - \mu(t) - \sum_{j=1}^{p} \xi_j \varphi_j(t) \right)^2 = 0.$$

- $\mathbb{E}(\xi_j) = 0$ *y* $\mathrm{Cov}(\xi_j, \xi_l) = \lambda_j \delta_{j,l}$ *para* $j, l \in \mathbb{N}$.

Demostración. Ver teorema 1.5 de Bosq (2000) [2].

De manera análoga a la esperanza, se puede definir un estimador para el núcleo de la covarianza como

$$\hat{c}_n(s,t) = \frac{1}{n} \sum_{j=1}^{n} (X_j(s) - \overline{X}_n(s))(X_j(t) - \overline{X}_n(t)).$$

Como en el caso finito-dimensional, no es insesgado pero sí lo es asintóticamente

Proposición 2.2.8. *Sean $(X_j)_{j\in\mathbb{N}}$, una sucesión de funcionales aleatorios en $L^2(I)$ de cuadrado integrable, i.i.d., entonces $\mathbb{E}(\hat{c}) = \frac{n}{n-1} c$.*

2.3. Representación de funciones usando bases

Dada una base de funciones $\{\phi_k\}_k$, una función $x(t)$ se puede aproximar bien en cierto sentido como $x(t) \approx \sum_{k=1}^{K} a_k \phi_k(t)$ para K suficientemente grande. Bases conocidas son, por ejemplo, la de monomios que se usan para construir series de potencias:

$$1, t, t^2, \ldots, t^k, \ldots$$

Otra alternativa es la base trigonométrica de Fourier.

$$1, \mathrm{sen}(\omega t), \cos(\omega t), \mathrm{sen}(2\omega t), \cos(2\omega t), \ldots, \mathrm{sen}(k\omega t), \cos(k\omega t), \ldots$$

Usualmente, el problema con el que se presenta la representación de una función es con el de los datos funcionales dados en ciertos instantes. Por ejemplo, al dato funcional $x_i(t)$ sólo se lo conoce en $t_0, t_1, \ldots, t_{p_i}$ con $p_i \gg 1$. Notemos que no tienen por qué ser

los mismos instantes que en los otros datos de la muestra. Una solución posible es usar interpolación mediante alguna base en cada dato funcional. Esto podría, sin embargo, generar un sobreajuste de datos por lo que el valor de K es un parámetro importante a la hora de aproximar a la función. Un K menor por otra parte implicaría un suavizado de la curva pero también aumentaría el error de aproximación.

Si bien, en el presente trabajo no se trabajará con datos funcionales dados en ciertos instantes, se rescatan otros beneficios del desarrollo de los datos funcionales en bases ya conocidas. Por ejemplo, además de regularizador el desarrollo como suma finita permite trabajar de alguna forma en dimensión finita y derivar fácilmente.

2.3.1. Base de Fourier

Esta base tiene la propiedad de ser periódica y el parámetro ω determina el período $2\pi/\omega$. Se suele ajustar ω según la longitud del intervalo donde están los datos funcionales. Se puede ver que forman una base ortonormal en $L^2(I)$ por lo que cualquier función allí se aproxima por este tipo de funciones. Además, observemos la simplicidad del cálculo de sus derivadas

$$\frac{d}{dt}\operatorname{sen}(k\omega t) = k\omega\cos(k\omega t)$$
$$\frac{d}{dt}\cos(k\omega t) = -k\omega\sin(k\omega t)$$

En consecuencia, si los coeficientes del desarrollo en serie de Fourier es de la forma

$$(c_0, c_1, c_2, c_3, c_4, \ldots),$$

entonces los de su derivada son

$$(0, \omega c_1, -\omega c_2, 2\omega c_3, -2\omega c_4, \ldots),$$

y los de la derivada segunda son

$$(0, -\omega^2 c_1, -\omega^2 c_2, -4\omega^2 c_3, -4\omega^2 c_4, \ldots).$$

Este desarrollo es muy útil en funciones sin características locales fuertes o donde la curvatura se mantiene en todo el intervalo. No es apropiado su uso en datos donde se sospecha que hay discontinuidades en la función o en sus derivadas.

Capítulo 3

Modelos lineales funcionales

Recordemos el modelo lineal clásico, en él se disponía de ciertas observaciones $x_1, \ldots, x_n \in \mathbb{R}^p$ y ciertas respuestas para cada observación $y_1, \ldots, y_n \in \mathbb{R}$ y se supone que la relación que hay entre ellas viene dada por un modelo de la forma

$$y_i = \beta^T x_i + \varepsilon_i = \langle \beta, x_i \rangle + \varepsilon_i, \quad i = 1, \ldots n. \tag{3.1}$$

donde los ε_i son variables aleatorias de esperanza cero y $\beta \in \mathbb{R}^p$ es un parámetro a determinar. A las variables x se las llama regresoras o covariables. Tradicionalmente, la primera covariable es constantemente 1, de modo que β_1 es lo que se conoce como el intercept. En el modelo funcional, algunas de estas variables son curvas. Los modelos lineales funcionales se pueden clasificar según cuál de estas variables representa una curva.

Modelo función a escalar:

$$Y_i = \int \beta(t) X_i(t) dt + \varepsilon_i.$$

En este caso, las variables regresoras y el parámetro desconocido son funciones y las respuestas son escalares. Si se asume que β y X_i están en L^2, se puede escribir como $Y_i = \langle \beta, X_i \rangle + \varepsilon_i$, quedando más explicita la generalización de este modelo respecto del de $\mathbb{R}^p$.

Modelo escalar a función:

$$Y_i(t) = \beta(t)^T X_i + \varepsilon_i(t).$$

Las variables regresoras representan escalares pero el parámetro desconocido β y las respuestas Y_i son curvas.

Modelo función a función:

$$Y_i(t) = \int \beta(t, s) X_i(s) ds + \varepsilon_i(t).$$

Las variables regresoras son funciones y las respuestas también. Acá β aparece representado por el núcleo de un operador integral.

La característica en común de estos modelos es que los parámtetros desconocidos son funcionales infinito-dimensionales y deben ser estimados por una muestra finita. Además, el nombre 'lineal' es consistente con el hecho de que las respuestas son el resultado de un operador lineal sobre los parámetros funcionales más un error. Si bien, un ajuste perfecto es posible, esto generaría un sobreajuste dando lugar a un mal modelo predictor. Se suele imponer entonces cierta suavidad en el modelo o también restringir los operadores involucrados a algún subespacio adecuado. En esta sección se revisará el primer tipo de modelo funcional. Previo a eso, se revisará el modelo lineal clásico.

3.1. Inferencia en el caso finito-dimensional

Asumiendo ε_i independientes y normales de esperanza nula y misma varianza σ^2 (homocedasticidad) se define el siguiente estimador de β para el modelo 3.1.

Definición 3.1.1. *Dadas las observaciones $x_1, \dots, x_n \in \mathbb{R}^p$ y sus respuestas $y_i = \beta^T x_i + \varepsilon_i$, $i = 1, \dots, n$, se define el estimador de mínimos cuadrados para β como aquel que minimiza la suma residual de cuadrados, es decir:*

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2.$$

El estimador de mínimos cuadrados para β se puede hallar resolviendo un sistema de ecuaciones:

Proposición 3.1.1. *Para el modelo lineal definido en 3.1, llamemos las matrices $\mathbf{X} \in \mathbb{R}^{n \times p}$ e $\mathbf{Y} \in \mathbb{R}^n$ como*

$$\mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

entonces el estimador de mínimos cuadrados es solución del sistema (ecuaciones normales)

$$\mathbf{X}^T \mathbf{X} \hat{\beta} = \mathbf{X}^T \mathbf{Y}.$$

En particular, si las observaciones son l.i. entonces $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$.

Reescribiendo al estimador como $\hat{\beta} = \beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\varepsilon}$, con $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$, se pueden deducir ciertos resultados teóricos sobre él.

Proposición 3.1.2. *Para el modelo lineal definido en 3.1 con observaciones l.i., si $\hat{\beta}$ es el estimador de mínimos cuadrados entonces*

- *$\hat{\beta}$ es insesgado,*
- *$\hat{\beta}$ es consistente y,*
- *$\hat{\beta} \sim N_p(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.*

La predicción $\hat{y}$ de una observación x se obtiene mediante $\hat{y} = \hat{\beta}^T x$. Luego $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$. Además, alguna coordenada de $\hat{\beta}$ significativamente distinta de cero es indicio de que alguna covariable no influye en la respuesta y que podría no tenerse en cuenta en el modelo. Se puede hacer un test de hipótesis sobre las coordenadas de β para determinar si son nulas o no.

A continuación, se define una matriz que vincula las respuestas con las predicciones.

Definición 3.1.2. *Dadas las observaciones $x_1, \ldots, x_n$ y sus respuestas $y_1, \ldots, y_n \in \mathbb{R}$, se define la matriz hat como aquella matriz $H \in \mathbb{R}^{n\times n}$ que satisface $\hat{\mathbf{Y}} = H\mathbf{Y}$.*

Observemos que esta definición no se restringe a regresión lineal múltiple sino a cualquiera cuyas variables de respuesta están en $\mathbb{R}$. En el caso del modelo 3.1 mediante estimador de mínimos cuadrados, se tiene que $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Se cumplen además varias propiedades.

Proposición 3.1.3. *Para el modelo lineal 3.1 con observaciones l.i., se tiene que*

- *H es simétrica e idempotente,*
- *$H\mathbf{X} = \mathbf{X}$,*
- *los autovalores de H pueden ser 0 o 1,*
- *$\mathrm{rg}(\mathbf{X}) = \mathrm{rg}(H) = \mathrm{tr}(H)$.*

De esta proposición se deduce que H es una matriz de proyección sobre el espacio columna de $\mathbf{X}$. Se interpreta entonces a $\hat{\mathbf{Y}}$ como la proyección ortogonal del vector $\mathbf{Y}$ sobre el espacio columna de $\mathbf{X}$.

3.2. Modelo función a escalar

Anteriormente se discutió la interpretación de las coordenadas de β cercanas a cero y la influencia de ciertas covariables en la respuesta. En el caso funcional se quiere que suceda alguna interpretación similar. En estas dimensiones lo que pasa es que los intervalos de valores t en donde $|\beta(t)|$ es grande las observaciones tienen gran peso allí.

El signo de β determinará si la asociación es positiva o negativa. Un estimador cuya curva asociada presenta saltos abruptos impide esta interpretación.

Un enfoque ingenuo para resolver el problema de hallar el parámetro funcional es transformar las ecuaciones normales a una versión funcional. O sea, las siguientes ecuaciones

$$\frac{1}{n}\mathbf{X}^T\mathbf{X}\hat{\beta} = \frac{1}{n}\mathbf{X}^T\mathbf{Y},$$

mirando el k-ésimo lugar de la ecuación,

$$\sum_{j=1}^{p}\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)_{kj}\hat{\beta}_j = \left(\frac{1}{n}\mathbf{X}^T\mathbf{Y}\right)_k,$$

se puede observar que $\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)_{kj} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{ik}\mathbf{X}_{ij}$ estima la esperanza del producto de la k-ésima y la j-ésima covariable, mientras que $\left(\frac{1}{n}\mathbf{X}^T\mathbf{Y}\right)_k = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{ik}\mathbf{Y}_i$ estima la esperanza de la k-ésima covariable con la variable respuesta Y. Definimos los operadores análogos a éstos en su versión funcional poblacional como

$$c_X(t,s) = \mathbb{E}(X(t)X(s)), \qquad\qquad c_{XY}(t) = \mathbb{E}(X(t)Y).$$

Se plantea entonces hallar β que cumpla la siguiente ecuación

$$\int c_X(t,s)\beta(s)ds = c_{XY}(t).$$

Un primer problema de esta aproximación es la dificultad de interpretar el resultado. Suponiendo que sólo se quiere conocer a β en una grilla y que este problema tiene solución, la función tendría mucho ruido. Esto en principio se debe a que no hay restricciones de suavidad

El segundo problema es que el operador integral $\varphi \longmapsto \int c_X(\cdot,s)\varphi(s)ds$ al ser compacto, no es inversible. Esto no asegura que el β se pueda encontrar bajo estas ecuaciones.

El objetivo de esta sección es brindar algunos métodos para determinar la curva β de este modelo con intercept α:

$$Y_i = \alpha + \int \beta(t)X_i(t)dt + \varepsilon_i, \quad i = 1,2,\ldots,n. \tag{3.2}$$

3.2.1. Estimación por bases

Una forma de encarar este problema es asumir que β admite un desarrollo como suma de funciones de una base determinística en L^2 de esta forma

$$\beta(t) = \sum_{k=1}^{K} c_k B_k(t),$$

donde K está fijo. Entonces se buscan $\alpha, c_1, \ldots, c_K$ que minimicen la suma residual de cuadrados

$$S(\alpha, c_1, \ldots, c_K) = \sum_{i=1}^{n} \left[Y_i - \alpha - \int \left(\sum_{k=1}^{K} c_k B_k(t) \right) X_i(t) dt \right]^2 . \tag{3.3}$$

La estructura de las funciones de la base influyen en la forma del estimador. Si a las observaciones X_i se las conoce en una grilla de t_j, usualmente el K se elige menor que la cantidad de nodos temporales para garantizar cierta suavidad en el estimador. Aunque hay métodos para elegir un K adecuado según el objetivo deseado, también es común experimentar con varios K. Se puede reescribir el modelo lineal funcional 3.2 en un modelo lineal como en 3.1:

$$Y_i = \alpha + \int \sum_{k=1}^{K} c_k B_k(t) X_i(t) dt + \varepsilon_i = \alpha + \sum_{k=1}^{K} c_k \int B_k(t) X_i(t) dt + \varepsilon_i.$$

Luego, definiendo $c, \tilde{X}_i \in \mathbb{R}^{K+1}$, $i = 1, 2, \ldots, n$ como

$$c = (\alpha, c_1, \ldots, c_K)^T \qquad \tilde{X}_i = \left(1, \int B_1(t) X_i(t) dt, \ldots, \int B_K(t) X_i(t) dt \right)^T , \tag{3.4}$$

se tiene que el modelo 3.2 y la suma 3.3 se reescriben como

$$Y_i = c^T \tilde{X}_i + \varepsilon_i \quad i = 1, 2, \ldots, n, \qquad\qquad S(c) = \sum_{i=1}^{n} (Y_i - c^T \tilde{X}_i)^2 \tag{3.5}$$

Por lo tanto, el problema de hallar el α y la curva β se redujo a buscar parámetros de un problema de modelo lineal finito-dimensional. Esto como resultado de considerar a β en el subespacio generado por $\{B_1, \ldots, B_K\}$. Entonces un estimador para c por mínimos cuadrados está dado por $\hat{c} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{Y}$ y los valores predichos se los calcula como $\hat{\mathbf{Y}} = \tilde{\mathbf{X}} \hat{c}$. De esto último se deduce que la matriz hat está dada por $H = \tilde{\mathbf{X}} (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T$

Respecto al sesgo del estimador, se observa que se restringió a β a un subespacio en L^2, mientras que en $\mathbb{R}^p$, no se imponía ninguna restricción sobre los parámetros. Sin embargo, se puede expandir sin ningún término de error usando una suma infinita y asumiendo que $\{B_k\}_k$ es completo. Es decir, β sin truncar queda $\beta(t) = \sum_{k=1}^{\infty} c_k B_k(t)$. Asumiendo un estimador del parámetro funcional con los primeros K términos de la base, entonces definimos el error de truncado como $\delta(t) := \sum_{k=K+1}^{\infty} c_k B_k(t)$. Entonces tenemos el siguientes desarrollo para β

$$\beta(t) = \sum_{k=1}^{K} c_k B_k(t) + \delta(t),$$

y para las respuestas usando 3.4,

$$Y_i = \alpha + \int \left(\sum_{k=1}^{K} c_k B_k(t) + \delta(t) \right) X_i(t) dt + \varepsilon_i = c^T \tilde{X}_i + \delta_i + \varepsilon_i,$$

donde $\delta_i = \int \delta(t)X_i(t)dt$. Luego el estimador para c se escribe como

$$\hat{c} = c + (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T(\boldsymbol{\delta} + \boldsymbol{\varepsilon}).$$

Fijado un K, entonces el estimador ahora está sesgado debido al término δ. Esto permite intuir que para obtener un estimador asintóticamente insesgado se precisa que K tienda a infinito a medida que n también lo haga.

3.2.2. Estimación con términos de penalidad

Desde este enfoque se busca imponer suavidad agregando un término adicional de penalización a la suma residual de cuadrados a minimizar en lugar de ajustar la cantidad de términos al expresar β. Se busca minimizar entonces

$$S_\lambda(\alpha, \beta) = \sum_{i=1}^{n} \left(Y_i - \alpha - \int \beta(t)X_i(t)dt \right)^2 + \lambda \int [D\beta(t)]^2 dt, \tag{3.6}$$

con $\lambda > 0$ y donde D es un operador diferencial aplicado sobre β y λ es el parámetro de suavizado. Es común elegir D como el operador derivada segunda: $D\beta = \beta''$. La idea de esta minimización consiste en forzar la suavidad al pedir que simultáneamente se minimice el término de penalización que involucra a la derivada $\lambda \int [D\beta(t)]^2 dt = \lambda ||D\beta||^2_{L^2}$. En este caso, el foco está en la elección del parámetro λ; si es muy grande, esto forzará a que β sea muy suave al punto de perder la estructura deseada mientras λ muy pequeño generarán β con mucho ruido y errores de forma aleatoria a lo largo del intervalo.

En la práctica, al igual que en el enfoque anterior, se empieza asumiendo a β en un subespacio de dimensión mucho mayor que la cantidad de nodos conocidos de los datos. Es decir, se asume $\beta(t) = \sum_{k=1}^{K} c_k B_k(t)$ con K muy grande tal que pequeñas variaciones de éste no afecten a β. Esto es para que el parámetro λ sea el que controle la suavidad y no K. Usando este desarrollo y 3.4, la expresión 3.6 se reescribe de esta forma:

$$\begin{aligned}
S_\lambda(\alpha, \beta) &= S_\lambda(c) \\
&= \sum_{i=1}^{n} \left(Y_i - \alpha - \sum_{k=1}^{K} c_k \int B_k(t)X_i(t)dt \right)^2 + \lambda \int \left[\sum_{k=1}^{K} c_k (DB_k)(t) \right]^2 dt \\
&= \sum_{i=1}^{n} (Y_i - c^T\tilde{X}_i)^2 + \lambda \sum_{k=1}^{K}\sum_{j=1}^{K} c_k c_j \int (DB_k)(t)(DB_j)(t)dt \\
&= (\mathbf{Y} - \tilde{\mathbf{X}}c)^T(\mathbf{Y} - \tilde{\mathbf{X}}c) + \lambda c^T\mathbf{R}c \\
&= c^T(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \lambda\mathbf{R})c - 2(\tilde{\mathbf{X}}^T\mathbf{Y})^T c + \mathbf{Y}^T\mathbf{Y},
\end{aligned}$$

donde $\mathbf{R}$ es una matriz de $(K+1)\times(K+1)$ dada por

$$\mathbf{R} := \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & \int(DB_1)(DB_1) & \int(DB_1)(DB_2) & \cdots & \int(DB_1)(DB_K) \\ 0 & \int(DB_2)(DB_1) & \int(DB_2)(DB_2) & \cdots & \int(DB_2)(DB_K) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \int(DB_K)(DB_1) & \int(DB_K)(DB_2) & \cdots & \int(DB_K)(DB_K) \end{pmatrix}.$$

Minimizando la función cuadrática resultante, el estimador del parámetro c está dado por

$$\hat{c} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \lambda\mathbf{R})^{-1}\tilde{\mathbf{X}}^T\mathbf{Y}.$$

Como sucedía con K en la sección anterior, es común inspeccionar diversos valores de λ y elegir aquel que parezca más adecuado. Otras formas de encarar esto es obtener un λ óptimo respecto a alguna propiedad sobre β o los datos.

3.2.3. Estimación por componentes principales funcionales

Asumiendo que los datos funcionales X_i provienen de un funcional aleatorio integrable en $L^2(I)$, entonces el funcional en cuestión admite un desarrollo de Karhunen-Loéve

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_j \varphi_j(t).$$

En este caso, nuevamente se usará una versión truncada. Recordando la proposición 2.2.7, ítem 3, mientras más términos se use, el truncado resulta más parecido al funcional original. Además, dado que los autovalores están ordenados de manera decreciente, con un p no necesariamente grande se obtiene una buena aproximación en el sentido de la norma $\mathbb{E}(||\cdot||^2)$.

Denotaremos $\hat{\mu}$ a la media muestral y $\hat{\varphi_j}$ a la estimación de la j-ésima componente principal funcional de X. Para un valor de p dado, se usa la siguiente aproximación para las observaciones

$$X_i(t) \approx \hat{\mu}(t) + \sum_{j=1}^{p} \hat{\xi}_{ij}\hat{\varphi}_j(t),$$

con $\hat{\xi}_{ij} = \int_I [X_i(t) - \hat{\mu}(t)]\hat{\varphi}_j(t)dt$. Luego se considera el siguiente modelo lineal aproximado en las observaciones:

$$\begin{aligned} Y_i &= \alpha + \int_I \beta(t)\left(\hat{\mu}(t) + \sum_{j=1}^{p}\hat{\xi}_{ij}\hat{\varphi}_j(t)\right)dt + \varepsilon_i \\ &= \alpha + \int_I \beta(t)\hat{\mu}(t)dt + \sum_{j=1}^{p}\hat{\xi}_{ij}\left(\int_I \beta(t)\hat{\varphi}_j(t)dt\right) + \varepsilon_i \\ &= b^T\boldsymbol{\xi}_i + \varepsilon_i, \end{aligned}$$

donde $b, \boldsymbol{\xi}_i \in \mathbb{R}^{p+1}$ para $i = 1, 2, \ldots, n$ están dados por

$$b = \left(\alpha + \int_I \beta(t)\hat{\mu}(t)dt, \int_I \beta(t)\hat{\varphi}_1(t)dt, \ldots, \int_I \beta(t)\hat{\varphi}_p(t)dt \right)^T \quad \text{y} \quad \boldsymbol{\xi}_i = (1, \hat{\xi}_{i1}, \ldots, \hat{\xi}_{ip})^T.$$

Se redujo entonces a un problema de regresión lineal en dimensión finita. Es importante observar que el objetivo ahora no es encontrar la curva β ni tampoco α, sino la construcción del modelo. Esto se debe a que no se usa directamente las observaciones sino sus coordenadas en las componentes principales funcionales.

Capítulo 4

Modelo lineal funcional truncado

En este capítulo se estudiará una modificación del modelo visto anteriormente que añade otros parámetros reales a determinar en los límites de integración, es decir que se trabajará con un modelo de la siguiente forma

$$Y_i = \alpha + \int_u^v \beta(t)X_i(t)dt + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{4.1}$$

donde además de α y β, hay que estimar $u, v \in \mathbb{R}$ donde $[u, v] \subsetneq I$. Más allá del nombre, el modelo deja de considerarse lineal pues en los parámetros ya no lo es.

Se puede interpretar que hay un período desconocido de tiempo contenido en I que es el relevante en los funcionales predictores a la hora de obtener las respuestas. Previo a mostrar los métodos para estimarlos, se tratará la buena definición del método.

4.1. Identificabilidad del modelo

Para que el modelo 4.1 esté "bien definido"necesitamos que no haya otro pero de diferentes parámetros que replique las mismas respuestas para los mismos datos. Esto es lo que define un modelo identificable. En nuestro caso, supongamos que existen otros parámetros $\alpha_1, \beta_1(t), u_1, v_1$ tal que $[u_1, v_1] \subsetneq I$ y satisfacen

$$P\left(\alpha + \int_u^v \beta(t)X(t)dt = \alpha_1 + \int_{u_1}^{v_1} \beta_1(t)X(t)dt\right) = 1,$$

el objetivo sería ver que $\alpha = \alpha_1$, $\beta = \beta_1$, $u = u_1$ y $v = v_1$. En el suplemento del artículo de Hall y Hooker [11], se demuestra la identificabilidad bajo las condiciones de X de cuadrado integrable y su operador de covarianza asociado de rango completo en $L^2(I)$.

4.2. Métodos para determinar los parámetros

Se analizará el caso $I = [0, l]$ y $u = 0$, es decir que sólo se intentará estimar un límite de integración que notaremos θ ($< l$) en lugar de v. Para β usaremos una expresión en una base de componentes principales funcionales estimadas $\{\varphi_j\}_j$ en el que omitiremos la notación con sombrero: $\beta = \sum_{j=1}^{m} \beta_j \varphi_j$. El valor de m, igual que antes, se interpreta como un parámetro de suavizado para la curva a estimar.

4.2.1. Método A: Inferencia simultánea

Suponemos ya conocidos $m \in \mathbb{N}$ y $\lambda > 0$, la intención es encontrar $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_m$ y $\hat{\theta}$ en un solo paso minimizando la siguiente expresión

$$S(\alpha, \beta_1, \ldots, \beta_m, \theta) = \sum_{i=1}^{n} \left[Y_i - \alpha - \int_0^{\theta} \left(\sum_{j=1}^{m} \beta_j \varphi_j(t) \right) X_i(t) dt \right]^2 .$$

En la práctica se minimiza esta expresión penalizada en su lugar para cada m

$$\tilde{S}(\alpha, \beta_1, \ldots, \beta_m, \theta) = \sum_{i=1}^{n} \left[Y_i - \alpha - \int_0^{\theta} \left(\sum_{j=1}^{m} \beta_j \varphi_j(t) \right) X_i(t) dt \right]^2 + n\lambda\theta^2$$

y se obtienen sus respectivos $\hat{\alpha}$ y $\hat{\beta}$.

El término de penalización ($n\theta^2$) se incluye para garantizar un subconjunto propio adecuado ya que minimizando la primera función objetivo se suele obtener estimadores cercanos a l. Más precisamente, un λ grande produce un estimador $\hat{\theta} \ll l$ mientras que uno pequeño da estimadores con los mismos problemas que el no penalizado. El orden de la penalización, n en este caso, es para que ambos términos sean comparables.

A continuación se propone una forma de encontrar λ y m previo a usar el método.

4.2.1.1. Obtención de parámetros λ y m

Para este método, se utilizará *general cross validation*, cuya descripción y detalle se encuentra en la sección 4.3.

Paso 1: Fijado $m = 3$, se obtiene λ usando *general cross validation.*

Paso 2: Para el λ obtenido y para cada m, se determinan $\hat{\alpha}$, $\hat{\beta}$ y $\hat{\theta}$ mediante el método A (notados $\hat{\alpha}^m$, $\hat{\beta}^m$ y $\hat{\theta}^m$).

Paso 3: Se elige aquel m que minimiza cierta función de pérdida $F(m)$. Algunas funciones propuestas para minimizar son:

1. Suma residual de cuadrados

$$F_1(m) = \sum_{i=1}^{n} \left[Y_i - \hat{\alpha}^m - \int_0^{\hat{\theta}^m} \hat{\beta}^m X_i(t) dt \right]^2 .$$

2. Con penalización en la derivada segunda

$$F_2(m) = \sqrt{F_1(m)} + \sqrt{\int_0^{\hat{\theta}^m} [D\hat{\beta}^m(t)]^2 dt}.$$

3. Penalizando con logaritmo

$$F_3(m) = \sqrt{F_1(m)} + \log\left(\int_0^{\hat{\theta}^m} [D\hat{\beta}^m(t)]^2 dt\right).$$

4.2.2. Método B: Inferencia por iteraciones

Suponemos ya conocidos $m \in \mathbb{N}$ y $\lambda > 0$. Primero se obtienen los estimadores $\check{\alpha}$ y $\check{\beta}$ para un modelo lineal funcional clásico como los de la sección 3.2., donde $\check{\beta}$ se encuentra generado por $\{\varphi_1, \ldots, \varphi_m\}$. Luego se halla θ minimizando la expresión

$$T(\theta) = \sum_{i=1}^{n} \left[Y_i - \check{\alpha} - \int_0^{\theta} \check{\beta}(t) X_i(t) dt\right]^2 + n\lambda\theta^2.$$

Finalmente definimos los estimadores a utilizar de la siguiente forma

- Truncando $\check{\beta}$ en el intervalo $[0, \theta]$

$$\hat{\beta}(t) = \begin{cases} \check{\beta}(t) & \text{si } t \leq \hat{\theta} \\ 0 & \text{si } t > \hat{\theta} \end{cases}.$$

- Definimos $\hat{\alpha}$ que corresponde al modelo truncado $Y = \hat{\alpha} + \int_0^{\hat{\theta}} \hat{\beta}(t)\overline{X}_n(t)dt$ a partir de la resta con el del modelo clásico $Y = \check{\alpha} + \int_0^{l} \check{\beta}(t)\overline{X}_n(t)dt$

$$\hat{\alpha} = \check{\alpha} - \int_{\hat{\theta}}^{l} \check{\beta}(t)\overline{X}_n(t)dt.$$

La forma de hallar λ difiere en algunos pasos con el del método anterior.

4.2.2.1. Obtención de parámetros λ y m

La descripción de *cross validation* utilizado en este método se encuentra la sección 4.3.

Paso 1: Para cada m, se determinan $\check{\alpha}$ y $\check{\beta}$ para el modelo funcional clásico en $[0, 1]$ (notados $\check{\alpha}^m$ y $\check{\beta}^m$).

Paso 2: Se elige m como en el paso 3 del método A efectuando los cambios correspondientes ($\hat{\alpha}^m$ por $\check{\alpha}^m$, $\hat{\beta}^m$ por $\check{\beta}^m$ y $\hat{\theta}^m$ por 1).

Paso 3: Para el m del paso anterior, se obtiene λ usando *cross validation*.

4.3. Cross validation

Dado un conjunto de datos $\{(x_1, y_1), \dots (x_n, y_n)\}$ y una propuesta de modelo $y = f(x) + \varepsilon$ que relaciona ambas variables, se desea estimar f con esos datos llamados *de entrenamiento*. Una vez obtenida la estimación $\hat{f}$, se desea evaluar el desempeño del método y modelo usado, por ejemplo, mediante el *error cuadrático medio* (abreviado ECM o MSE en inglés) dado por

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2.$$

El ECM será pequeño si los valores de las predicciones $\hat{f}(x_i)$ son cercanas a los valores verdaderos de las respuestas y_i y será grande si alguno de ellos difiere de manera significativa.

Si el objetivo del modelo es predecir, entonces no es interesante el ECM anterior pues éste sólo mide la calidad de la predicción con los datos de entrenamiento. Lo que se busca es elegir un método que logre una buena predicción con cualquier dato (x_j, y_j). Si se tiene un nuevo conjunto de datos $\{(x_1^*, y_1^*), \dots, (x_{\tilde{n}}^*, y_{\tilde{n}}^*)\}$, se puede medir la calidad de la predicción con el error cuadrático medio pero usando estos datos:

$$\text{MSPE} = \frac{1}{\tilde{n}}\sum_{j=1}^{\tilde{n}}(y_j^* - \hat{f}(x_j^*))^2.$$

Ese conjunto de datos recibe el nombre de *datos de validación*.

En ausencia de este último tipo de datos, se puede seleccionar modelos con buen nivel predictivo usando los datos de entrenamiento. Por ejemplo mediante el criterio de *cross validation* (validación cruzada). Éste consiste en dividir al conjunto de datos en dos partes: un conjunto de entrenamiento y otro de validación. Con el primer conjunto se estima $\hat{f}$ y con el segundo se evalúa el desempeño mediante la estimación del MSPE. A continuación se presentan dos variantes conocidas de este método.

4.3.1. Leave-one-out cross validation (LOOCV)

En esta versión, en lugar de crear dos subconjuntos de observaciones de tamaño comparable el conjunto de validación va a estar compuesta solamente por (x_1, y_1) mientras que el resto de las observaciones forman el conjunto de entrenamiento usado para estimar f (llamado $\hat{f}^{(-1)}$). Luego el error cuadrático medio de predicción para esta partición está dado por $\text{MSPE}_1 = (y_1 - \hat{f}^{(-1)}(x_1))^2$. Sin embargo esta estimación del MSPE no es buena pues está construida con una observación resultando muy variable.

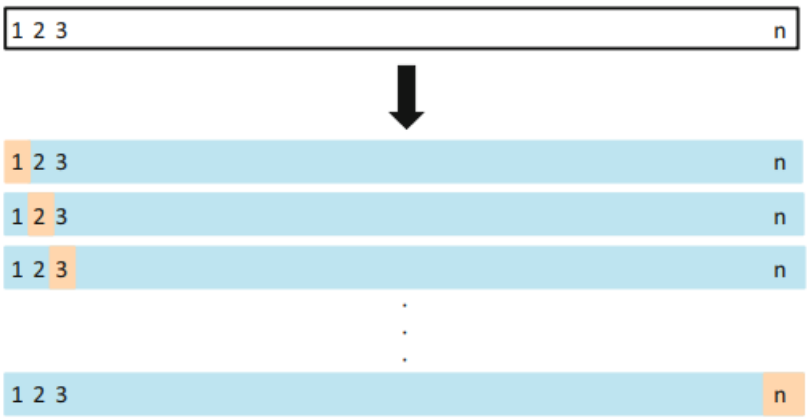


Figura 4.1: Una representación visual de LOOCV. Un conjunto de n datos es dividido en entrenamiento (en azul) y validación (anaranjado) n veces. El primer conjunto de entrenamiento tiene a todas la observaciones salvo la 1, segundo tiene a todas salvo la 2 y así. Luego se promedia los MSPE asociados a cada partición.

Se puede repetir este proceso seleccionando (x_2, y_2) como conjunto de validación y a las demás $n-1$ observaciones como conjunto de entrenamiento para obtener $\hat{f}^{(-2)}$ y computar $\text{MSPE}_2 = (y_2 - \hat{f}^{(-2)}(x_2))^2$. Repitiendo este enfoque con las demás observaciones se obtienen las demás estimaciones del error cuadrático medio de predicción: $\text{MSPE}_3, \ldots, \text{MSPE}_n$. En la figura 4.1 se ilustra la partición de los datos recién citada. Luego, la estimación del error de predicción para el LOOCV se obtiene como la media de las estimaciones de errores anteriormente calculados:

$$\text{LOOCV} = \frac{1}{n}\sum_{i=1}^{n}\text{MSPE}_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}^{(-i)}(x_i))^2.$$

El pseudocódigo para LOOCV es el siguiente:

Algorithm 1 Leave-one-out cross validation

Input: $(x_1, y_1), \ldots, (x_n, y_n)$
Output: $valError$
  $valError \leftarrow 0$
  $datos \leftarrow \{(x_1, y_1), \ldots, (x_n, y_n)\}$
  for i in 1 to n **do**
    $datoMenos \leftarrow datos - \{(x_i, y_i)\}$
    $\hat{f} \leftarrow \text{ajuste}(f, datoMenos)$
    $y_{pred} \leftarrow \hat{f}(x_i)$
    $valError \leftarrow valError + (y_i - y_{pred})^2$
  end for
  $valError \leftarrow \frac{valError}{n}$

LOOCV tiene la desventaja de ser potencialmente costoso ya que hay que realizar n estimaciones de $\hat{f}$. Esto puede consumir mucho tiempo si n es muy grande y si es lento para predecir la respuesta para un individuo. Este es uno de los motivos por los cuales se suele usar otra versión de cross validation: *k-Fold cross validation*.

4.3.2. k-Fold cross validation

Este enfoque involucra dividir al conjunto de datos en k grupos de tamaño similar de forma aleatoria. El primer grupo es tratado como conjunto de validación y los datos de los grupos restantes son usados como entrenamiento para obtener una estimación de f. El error cuadrático medio de predicción MSPE_1 se obtiene usando los datos del conjunto de validación (el primer grupo). Este procedimiento se repite otras $k-1$ veces, cada una usando un grupo diferente como conjunto de validación como se ilusta en la figura 4.2. Esto resulta en k estimaciones del error de predicción: $\text{MSPE}_1, \ldots, \text{MSPE}_k$. La estimación del error para el k-fold CV se computa promediando esos valores:

$$\text{CV}_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \text{MSPE}_i.$$

A continuación, se detalla el pseudocódigo para k-fold CV:

Algorithm 2 k-Fold cross validation

Input: $(x_1, y_1), \ldots, (x_n, y_n), k$
Output: $valError$
  $valError \leftarrow 0$
  $datos \leftarrow \{(x_1, y_1), \ldots, (x_n, y_n)\}$
  $grupo_1, \ldots, grupo_k \leftarrow \text{particion}(datos, k)$
  for i in 1 to k **do**
    $error \leftarrow 0$
    $size \leftarrow \#grupo_i$
    $datoMenos \leftarrow datos - grupo_i$
    $\hat{f} \leftarrow \text{ajuste}(f, datoMenos)$
    for (x, y) in $grupo_i$ **do**
      $y_{pred} \leftarrow \hat{f}(x)$
      $error \leftarrow error + (y - y_{pred})^2$
    end for
    $valError \leftarrow valError + \frac{error}{size}$
  end for
  $valError \leftarrow \frac{valError}{k}$

Se puede observar que LOOCV es un caso particular de k-fold CV en el que k es igual a n. En la práctica, se suele usar $k = 5$ o $k = 10$. El propósito es puramente

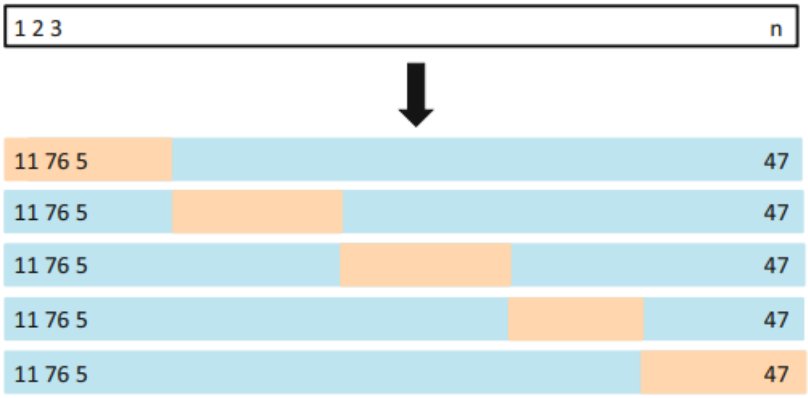


Figura 4.2: Una representación visual de 5-fold CV. Un conjunto de n datos es dividido en 5 subconjuntos disjuntos de forma aleatoria. En cada una de las 5 instancias, uno funciona como conjunto de validación (anaranjado) y el resto (azul), de entrenamiento. Luego se promedia los MSPE asociados a cada partición.

computacional pues se involucra k ajuste de datos, esto significa que para k mayores, requiere más tiempo para ejecutar el proceso.

Analizando el sesgo del error de predicción obtenido por este método, si n es grande se puede observar que para k cercanos a n se tienen estimaciones con poco sesgo para el error. Esto se debe a que se realizan tantos entrenamientos y ajustes como k con conjuntos de tamaño comparable con la muestra total. Por otra parte, k menores generan estimaciones con mayor sesgo. Entonces, desde el punto de vista del sesgo es preferible LOOCV en lugar de k-fold CV con $k < n$.

Respecto a la varianza, la situación es la inversa, para k cercanos a n lo que se hace en cada iteración es calcular errores de predicción con muestras con muchas observaciones en común resultando en estimaciones con mucha variabilidad.

4.3.3. General cross validation

El proceso de cross validation requiere más de un ajuste de datos y predicción con el resto. Si se requiere una estimación del error insesgado, hay que realizar leave-on-out cross validation que resulta poco eficiente con una gran cantidad de observaciones. General cross validation (Craven y Wahba, 2002) pretende aproximar los cálculos obtenidos por LOOCV sin tener que volver a estimar f a medida que varían los datos de entrenamiento. La estimación del error viene dada por la fórmula:

$$\text{GCV} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}{(1 - \text{tr}(H)/n)^2},$$

donde $H \in \mathbb{R}^{n\times n}$ es la matriz hat del modelo.

Previo a mostrar cómo este método generaliza a la versión leave-one-out de cross validation, se demostrará el siguiente resultado.

Lema 4.3.1. *Dado el modelo $y = f(x)+\varepsilon$ con $f \in F$ donde $\mathcal{F}$ es una familia de funciones y sean $\hat{f}$ y $\tilde{f}^{(-k)}$ los estimadores de mínimos cuadrados de f en $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ y $\{(x_1, y_1), \ldots, (x_{k-1}, y_{k-1}), (x_k, \hat{f}^{(-k)}(x_k)), (x_{k+1}, y_{k+1}), \ldots, (x_n, y_n)\}$, entonces $\tilde{f}^{(-k)} = \hat{f}^{(-k)}$.*

Demostración. Por definición, hay que ver que $\hat{f}^{(-k)}$ resuelve el siguiente problema de minimización

$$\min_{f\in\mathcal{F}} \frac{1}{n}\left[\sum_{\substack{i=1\\ i\neq k}}^{n}(y_i - f(x_i))^2 + (\hat{f}^{(-k)}(x_k) - f(x_k))^2\right]$$

Sea $f \in \mathcal{F}$, entonces

$$\frac{1}{n}\left[\sum_{\substack{i=1\\ i\neq k}}^{n}(y_i - \hat{f}^{(-k)}(x_i))^2 + (\hat{f}^{(-k)}(x_k) - \hat{f}^{(-k)}(x_k))^2\right] = \frac{1}{n}\sum_{\substack{i=1\\ i\neq k}}^{n}(y_i - \hat{f}^{(-k)}(x_i))^2$$

$$\leq \frac{1}{n}\sum_{\substack{i=1\\ i\neq k}}^{n}(y_i - f(x_i))^2 \leq \frac{1}{n}\left[\sum_{\substack{i=1\\ i\neq k}}^{n}(y_i - f(x_i))^2 + (\hat{f}^{(-k)}(x_k) - f(x_k))^2\right],$$

donde la primera desigualdad proviene de la definición del estimador $\hat{f}^{(-k)}$. Por lo tanto, se tiene que $\hat{f}^{(-k)}$ minimiza la expresión en $\mathcal{F}$. □

El siguiente resultado muestra una fórmula alternativa para GCV que se asemeja a la de LOOCV.

Proposición 4.3.2. *Dada una muestra $x_1, \ldots, x_n$ y sus respuestas $y_i = f(x_i) + \varepsilon_i$, $i = 1, \ldots, n$, se tiene que*

$$GCV = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}^{(-i)}(x_i))^2 w_i,$$

donde $w_i = \left[\dfrac{1 - h_{ii}}{1 - \mathrm{tr}(H)/n}\right]^2$ y H es la matriz hat.

Demostración. Por propiedad de estimadores de mínimos cuadrados en modelos lineales, se tiene que $(\hat{f}(x_1), \ldots, \hat{f}(x_n))^T = H\mathbf{Y}$ y $(\tilde{f}^{(-i)}(x_1), \ldots, \tilde{f}^{(-i)}(x_n))^T = H\mathbf{Y}^{(i)}$ donde $\mathbf{Y}^{(i)} = (y_1, \ldots, y_{i-1}, \hat{f}^{(-i)}(x_i), y_{i+1}, \ldots, y_n)^T$. Se tiene entonces la siguiente igualdad para cada $i = 1, \ldots, n$:

$$\hat{f}(x_i) - \hat{f}^{(-i)}(x_i) = \hat{f}(x_i) - \tilde{f}^{(-i)}(x_i) = \sum_{j=1}^{n} h_{ij}y_j - \sum_{j=1}^{n} h_{ij}y_j^{(i)} = h_{ii}(y_i - \hat{f}^{(-i)}(x_i)),$$

donde la primera igualdad vale por el lema anterior. Sumando y restando y_i al primer miembro se tiene lo siguiente

$$-(y_i - \hat{f}(x_i)) + (y_i - \hat{f}^{(-i)}(x_i)) = h_{ii}(y_i - \hat{f}^{(-i)}(x_i))$$
$$(1 - h_{ii})(y_i - \hat{f}^{(-i)}(x_i)) = y_i - \hat{f}(x_i).$$

Luego se reescribe la estimación de error según GCV:

$$\text{GCV} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}{(1 - \text{tr}(H)/n)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}[(y_i - \hat{f}^{(-i)}(x_i))(1 - h_{ii})]^2}{(1 - \text{tr}(H)/n)^2}$$
$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}^{(-i)}(x_i))^2 w_i.$$

□

Se puede observar que si los elementos de la diagonal de H son iguales, entonces LOOCV y GCV coinciden. En el caso del modelo lineal clásico, h_{ii} y $\text{tr}(H)/n$ distan a lo sumo en 1. Craven y Wahba mostraron que en el caso con f en el espacio de funciones periódicas de $W^{m,2}$ de integral cero, LOOCV y GCV son iguales.

Debido a la hipótesis de la existencia de la matriz hat, no siempre se puede usar GCV para seleccionar modelos. Incluso, su obtención podría ser costosa. En este trabajo, se puede obtener H para el método A y así usar este criterio para seleccionar un λ adecuado.

Proposición 4.3.3. *El método A para el modelo lineal truncado admite una matriz hat.*

Demostración. Si se define para cada $\theta \in I$, la siguente función

$$g_\theta(\alpha, \beta_1, \ldots, \beta_m) = \tilde{S}(\alpha, \beta_1, \ldots, \beta_m, \theta),$$

se tiene la igualdad

$$\min_{\substack{\theta \in I \\ \alpha,\beta_1,\ldots,\beta_m \in \mathbb{R}}} \tilde{S}(\alpha, \beta_1, \ldots, \beta_m, \theta) = \min_{\theta \in I} \min_{\alpha,\beta_1,\ldots,\beta_m \in \mathbb{R}} g_\theta(\alpha, \beta_1, \ldots, \beta_m).$$

Sea $\theta \in I$, minimizar g_θ es equivalente a minimizar la suma residual de cuadrados

$$S_\theta(\alpha, \beta_1, \ldots, \beta_m) = \sum_{i=1}^{n}\left[Y_i - \alpha - \int_0^\theta \left(\sum_{j=1}^{m}\beta_j\varphi_j(t)\right) X_i(t)dt\right]^2,$$

que se logra en $\hat{\beta}_\theta = (\tilde{\mathbf{X}}_\theta^T\tilde{\mathbf{X}}_\theta)^{-1}\tilde{\mathbf{X}}_\theta^T\mathbf{Y}$ donde para cada $i \in \{1, \ldots, n\}$, la fila i de $\tilde{\mathbf{X}}_\theta$ está dada por $\left(1, \int_0^\theta \varphi_1(t)X_i(t)dt, \ldots, \int_0^\theta \varphi_m(t)X_i(t)dt\right)$.

Sea $\theta^* \in I$ la que minimiza la expresión

$$\min_{\alpha,\beta_1,\ldots,\beta_m \in \mathbb{R}} g_\theta(\alpha, \beta_1, \ldots, \beta_m) = g_\theta(\hat{\beta}_\theta) = \tilde{S}(\hat{\beta}_\theta, \theta),$$

entonces $\tilde{S}$ se minimiza en $(\hat{\beta_{\theta^*}}, \theta^*)$ y la predicción se calcula como $\hat{\mathbf{Y}} = \tilde{\mathbf{X}}_\theta \hat{\beta_{\theta^*}} = H_{\theta^*}\mathbf{Y}$ con $H_{\theta^*} = \tilde{\mathbf{X}}_\theta(\tilde{\mathbf{X}}_\theta^T \tilde{\mathbf{X}}_\theta)^{-1}\tilde{\mathbf{X}}_\theta^T$. $\square$

Capítulo 5

Simulaciones y resultados

En este capítulo se realizará un análisis del desempeño de cada método. Para la base trigonométrica de norma 1 (en $L^2([0,1])$) dada en la sección 2.3 con $\omega = 2\pi$ y de 25 funciones, que se notarán $\psi_1, \ldots, \psi_{25}$, se generarán covariables $X_i \in L^2([0,1])$ para $i = 1, \ldots, 100$ a partir de ella como combinación lineal de sus elementos. Para cada i, se sortean de la siguiente forma

$$X_i = \sum_{k=1}^{25} a_k^i \psi_k,$$

con $a_k^i \sim N(0, \exp\{-(k-1)/4\})$ para cada k. Esto produce una menor presencia de términos de alta frecuencia. Una vez fijados α, β y θ, las variables de respuesta están dadas por

$$Y_i = \alpha + \int_0^\theta \beta(t) X_i(t) dt + \varepsilon_i,$$

con $\varepsilon_i \sim N(0,1)$.

Se analizará la calidad de las estimaciones y predicciones para los siguiente seis modelos dados por $\alpha = 0$, $\theta = 0{,}5$ y las siguientes funciones β:

- $\beta_1(t) = 100\psi_1(t)\mathbf{1}_{[0,\theta]}(t)$,
- $\beta_2(t) = 70\psi_2(t)\mathbf{1}_{[0,\theta]}(t)$,
- $\beta_3(t) = 40\big(\psi_1(t) + \psi_3(t)\big)\mathbf{1}_{[0,\theta]}(t)$,
- $\beta_4(t) = 25\big(2\psi_2(t) + \psi_4(t)\big)\mathbf{1}_{[0,\theta]}(t)$,
- $\beta_5(t) = 10\big(5\psi_2(t) + 4\psi_4(t) + \psi_6(t)\big)\mathbf{1}_{[0,\theta]}(t)$,
- $\beta_6(t) = 7\big(\psi_6(t) - 10\psi_{15}(t)\big)\mathbf{1}_{[0,\theta]}(t)$.

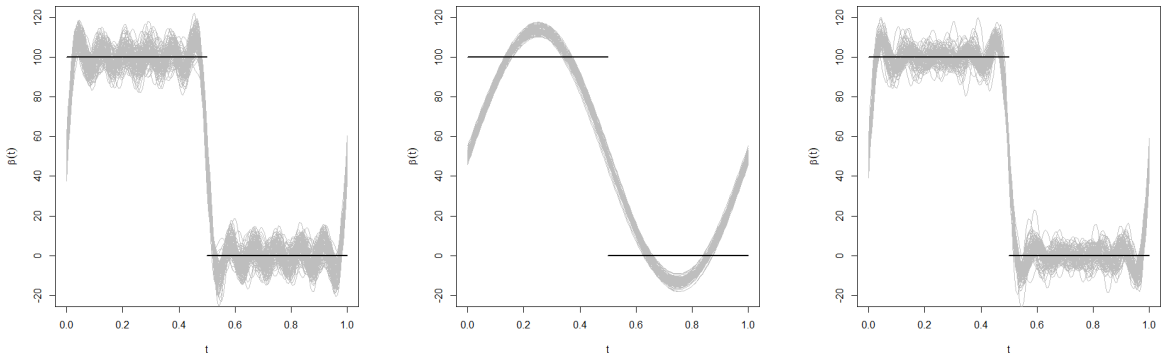
Se estudiará el comportamiento de los siguientes parámetros sobre la base de 100 replicaciones. En cada una de las simulaciones se calculará el MSE (error cuadrático medio) y el ISE (error cuadrático integral) para β, dado por $\int_I (\beta(t) - \hat{\beta}(t))^2 dt$.

5.1. Primer modelo

Una particularidad de la función β para este modelo es que hay una discontinuidad esencial en θ sin embargo las derivadas presentan una discontinuidad evitable allí.

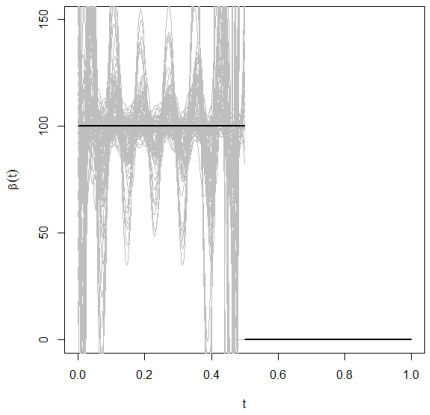
Método	Media $\hat{\theta}$	Desv. $\hat{\theta}$	MSE	ISE para β
A (con F_1)	0,487	0,009	$0,851$	$3\,810,515$
A (con F_2)	0,500	0	0,957	0,078
A (con F_3)	0,500	0	0,993	0,084
B (con F_1)	0,514	0,006	1,765	70,812
B (con F_2)	0,555	0,092	61,721	311,574
B (con F_3)	0,514	0,005	2,040	69,188
Sin truncado (con F_1)	-	-	0,740	107,022
Sin truncado (con F_2)	-	-	71,989	476,605
Sin truncado (con F_3)	-	-	0,840	106,670

Cuadro 5.1: Estimaciones de parámetros para el modelo con $\beta = \beta_1$.

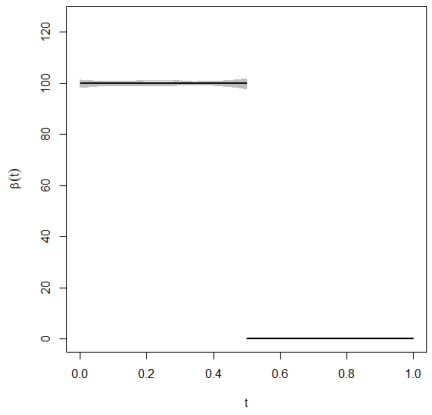


(a) Sin truncado con F_1. (b) Sin truncado con F_2. (c) Sin truncado con F_3.

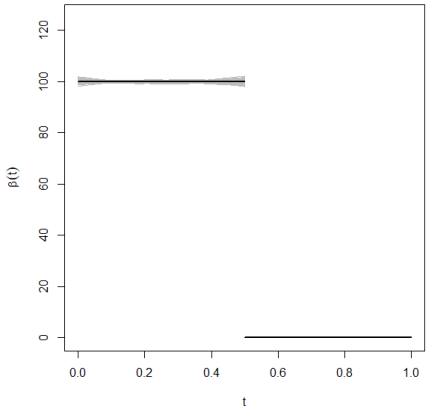
Figura 5.1: La función β está en negro y las estimaciones, en gris.



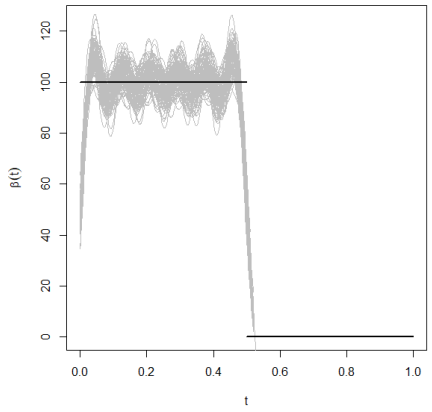
(a) Método A con F_1.



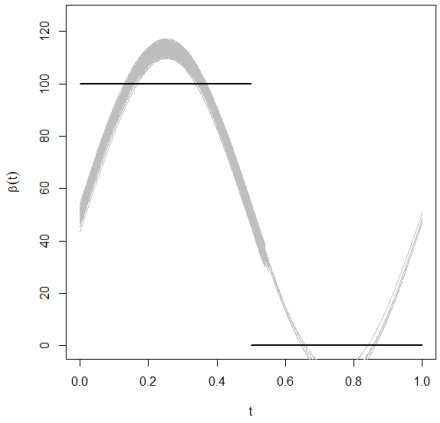
(b) Método A con F_2.



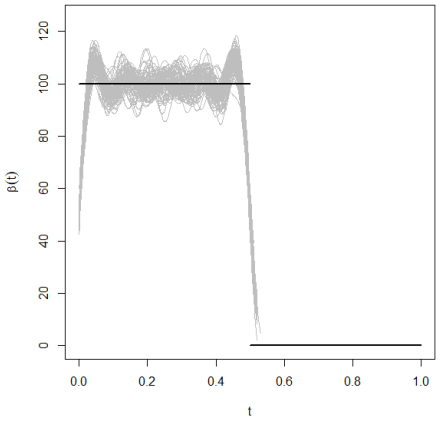
(c) Método A con F_3.



(d) Método B con F_1.

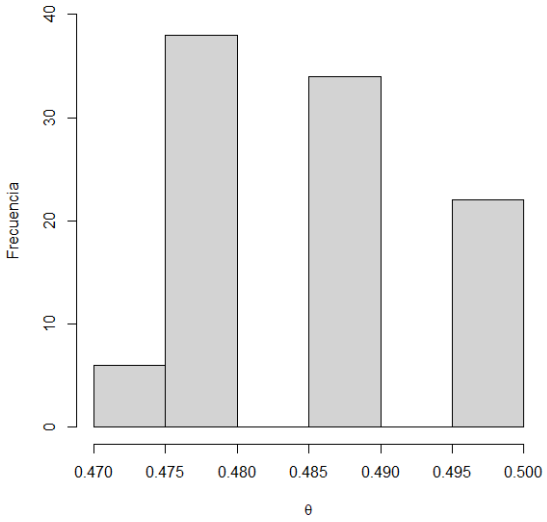


(e) Método B con F_2.

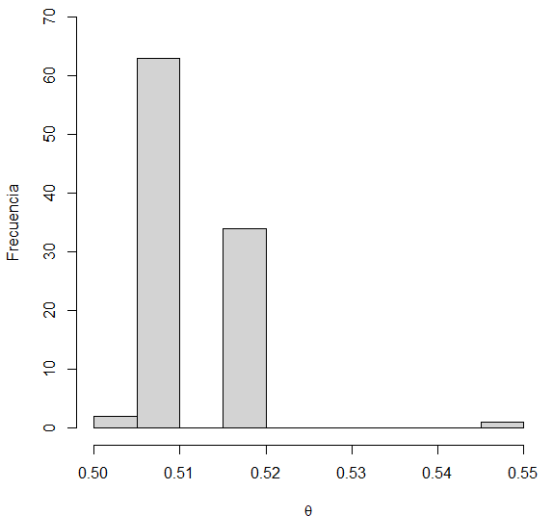


(f) Método B con F_3.

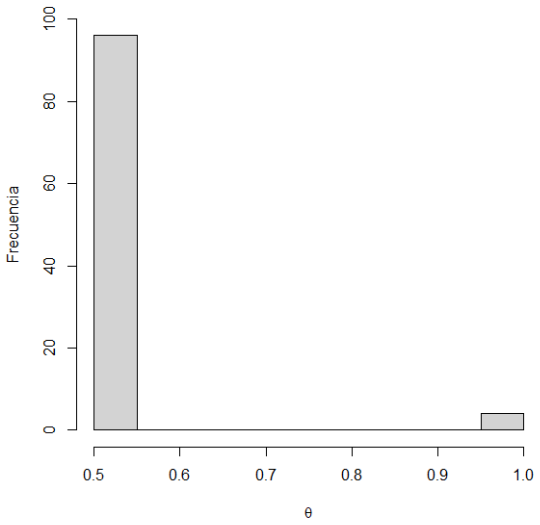
Figura 5.2: La función β está en negro y las estimaciones, en gris.



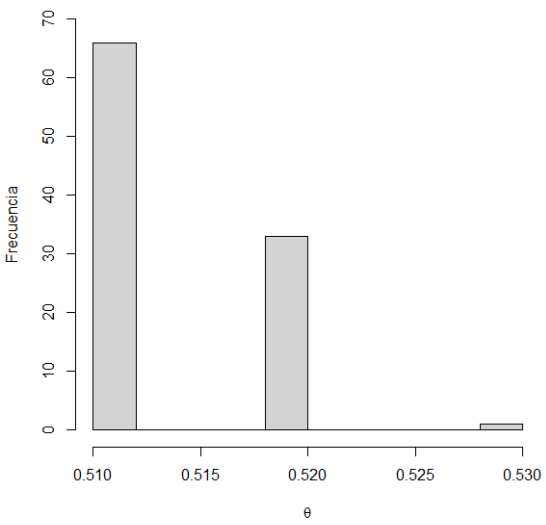
(a) Método A con F_1.



(b) Método B con F_1.



(c) Método B con F_2.



(d) Método B con F_3.

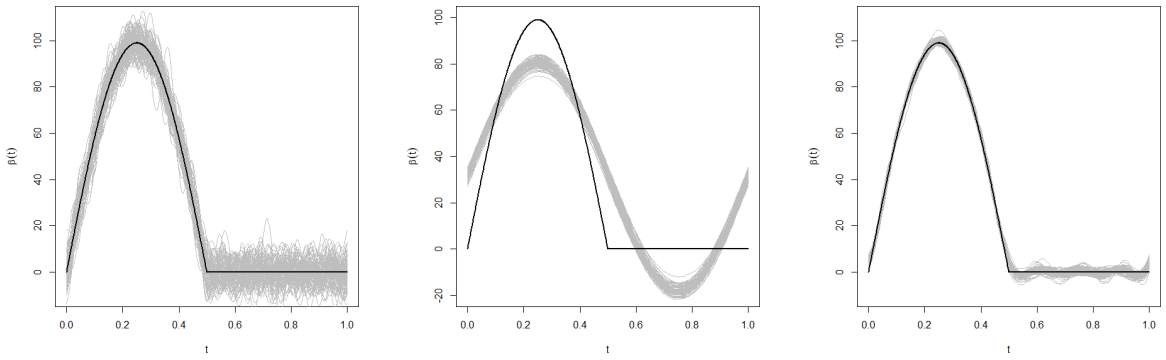
Figura 5.3: Histogramas para $\hat{\theta}$.

Para el método A, tanto con la función F_2 como con F_3 el valor de $\hat{\theta}$ siempre fue el mismo en cada una de las 100 simulaciones. Esto puede deberse a que para minimizar, se consideró θ en el espacio discreto $\{0, 0{,}01, 0{,}02, \ldots, 0{,}99, 1\}$. Por esta razón no fue necesario realizar sus histogramas.

5.2. Segundo modelo

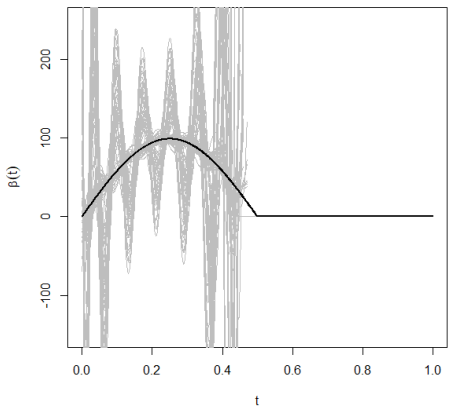
La función β en este modelo es continua en θ, no así su derivada que presenta una discontinuidad esencial. La derivada segunda presenta, en cambio, una discontinuidad evitable en θ.

Método	Media $\hat{\theta}$	Desv. $\hat{\theta}$	MSE	ISE para β
A (con F_1)	0,449	0,012	0,873	14 373,864
A (con F_2)	0,477	0,005	0,964	2,450
A (con F_3)	0,477	0,006	0,934	2,617
B (con F_1)	0,486	0,010	0,942	11,676
B (con F_2)	0,467	0,008	34,294	101,024
B (con F_3)	0,489	0,005	0,949	1,203
Sin truncado (con F_1)	-	-	0,747	23,853
Sin truncado (con F_2)	-	-	79,780	235,142
Sin truncado (con F_3)	-	-	0,893	2,273

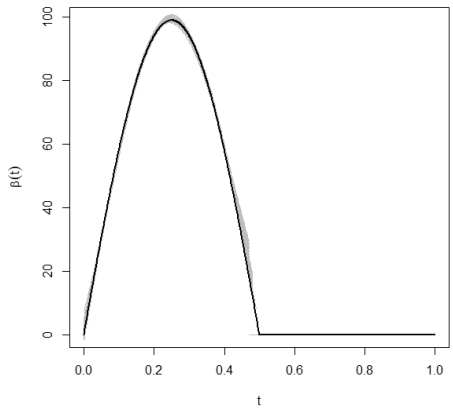


(a) Sin truncado con F_1. (b) Sin truncado con F_2. (c) Sin truncado con F_3.

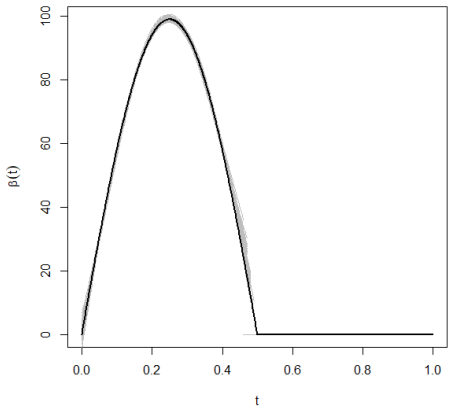
Figura 5.4: La función β está en negro y las estimaciones, en gris.



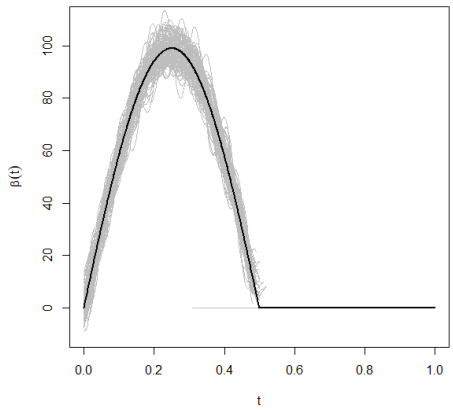
(a) Método A con F_1.



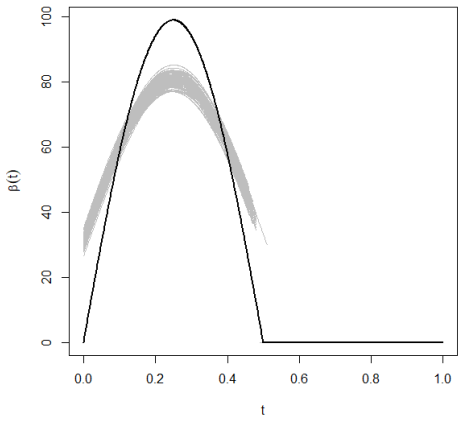
(b) Método A con F_2.



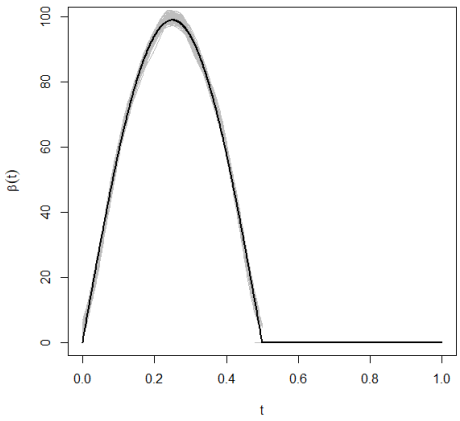
(c) Método A con F_3.



(d) Método B con F_1.

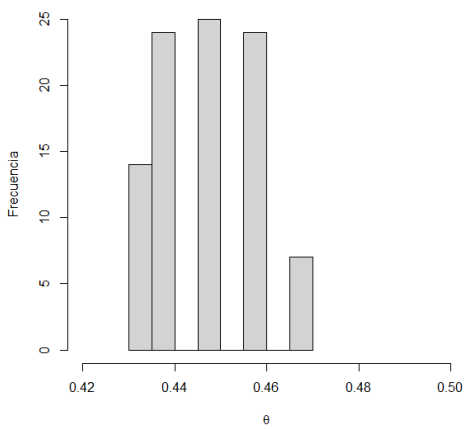


(e) Método B con F_2.

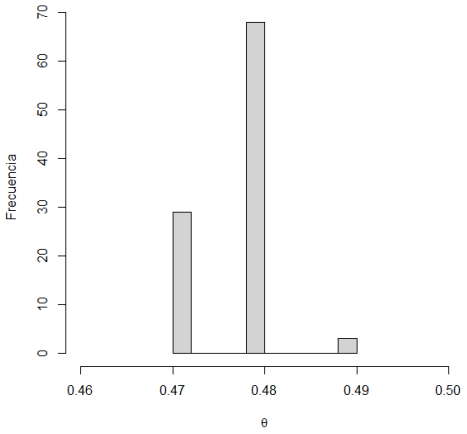


(f) Método B con F_3.

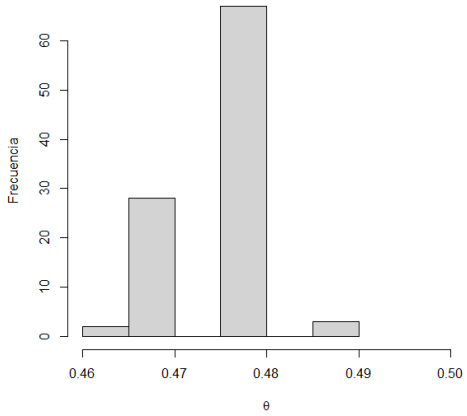
Figura 5.5: La función β está en negro y las estimaciones, en gris.



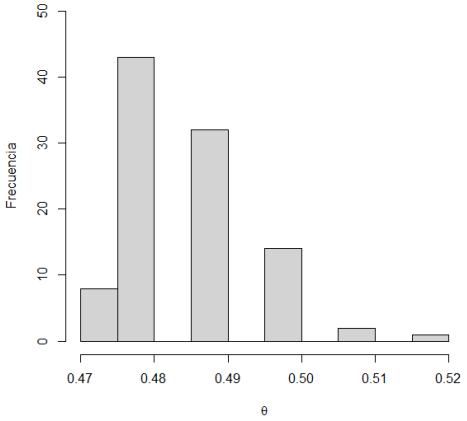
(a) Método A con F_1.



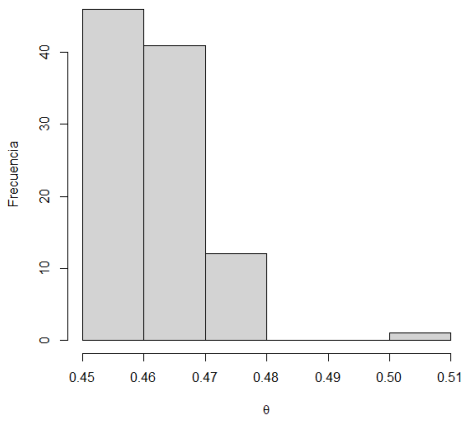
(b) Método A con F_2.



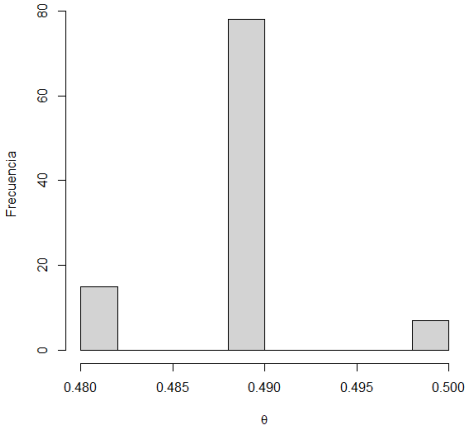
(c) Método A con F_3.



(d) Método B con F_1.



(e) Método B con F_2.



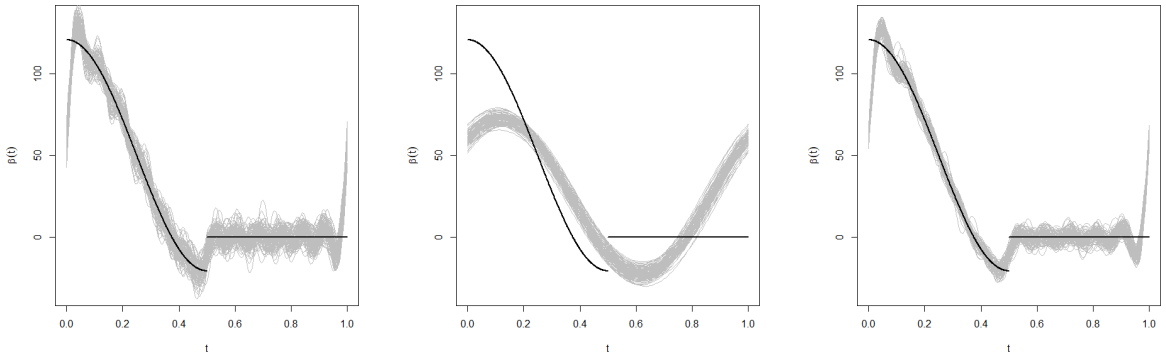
(f) Método B con F_3.

Figura 5.6: Histogramas para $\hat{\theta}$.

5.3. Tercer modelo

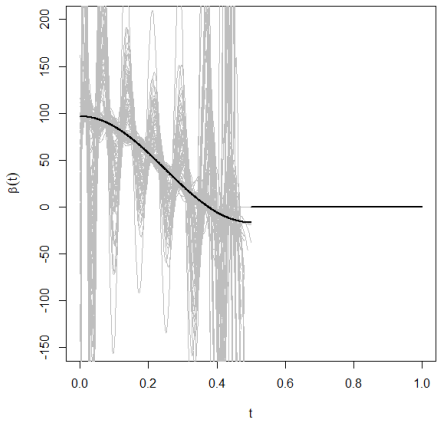
La función β presenta una discontinuidad esencial en θ, lo mismo con su derivada segunda. En cambio, en su derivada hay una discontinuidad evitable en θ.

Método	Media $\hat{\theta}$	Desv. $\hat{\theta}$	MSE	ISE para β
A (con F_1)	0,462	0,016	0,875	8 315,386
A (con F_2)	0,491	0,006	0,950	3,134
A (con F_3)	0,492	0,006	0,960	2,817
B (con F_1)	0,495	0,009	1,625	33,770
B (con F_2)	0,320	0,027	95,279	217,952
B (con F_3)	0,499	0,010	2,141	28,388
Sin truncado (con F_1)	-	-	0,739	85,271
Sin truncado (con F_2)	-	-	239,512	753,526
Sin truncado (con F_3)	-	-	0,839	82,361

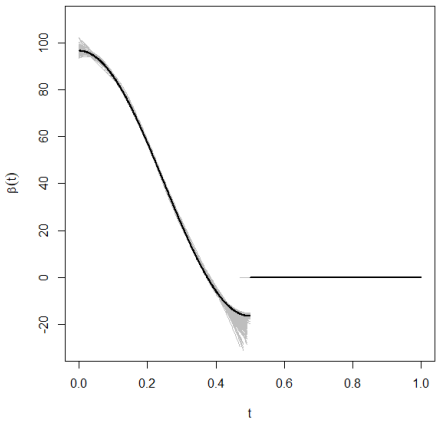


(a) Sin truncado con F_1. (b) Sin truncado con F_2. (c) Sin truncado con F_3.

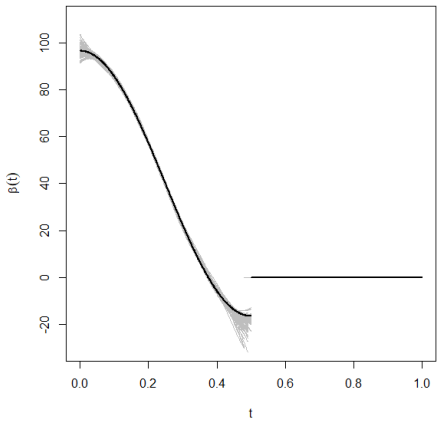
Figura 5.7: La función β está en negro y las estimaciones, en gris.



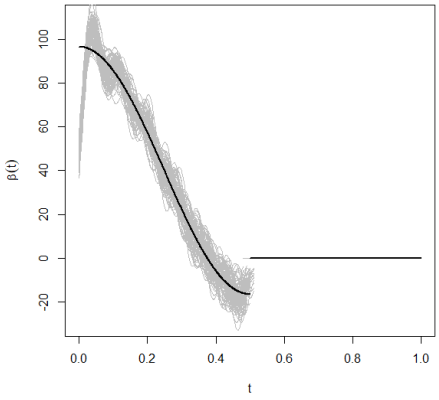
(a) Método A con F_1.



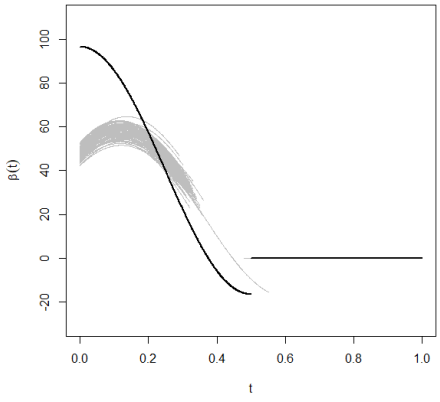
(b) Método A con F_2.



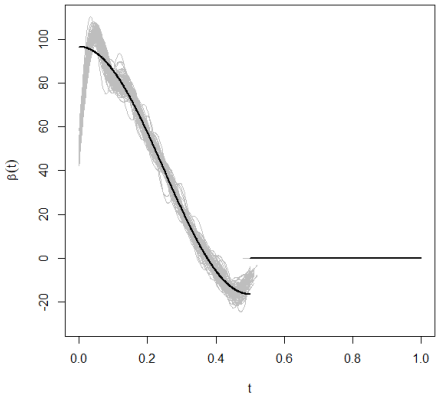
(c) Método A con F_3.



(d) Método B con F_1.

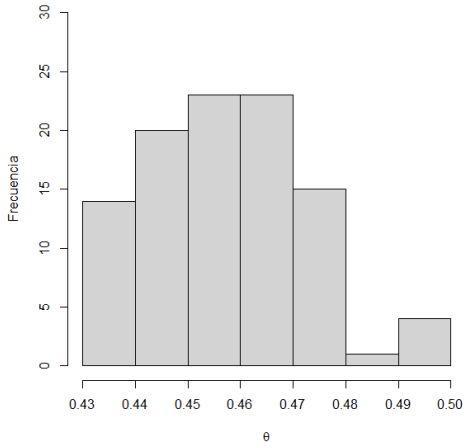


(e) Método B con F_2.

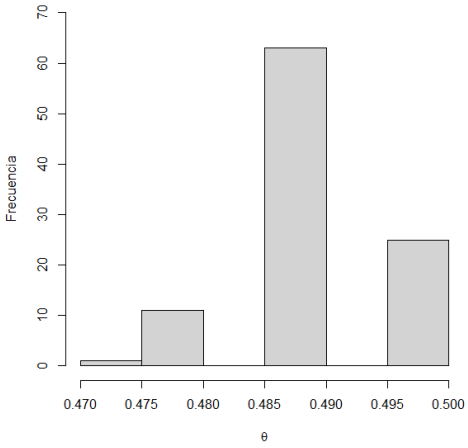


(f) Método B con F_3.

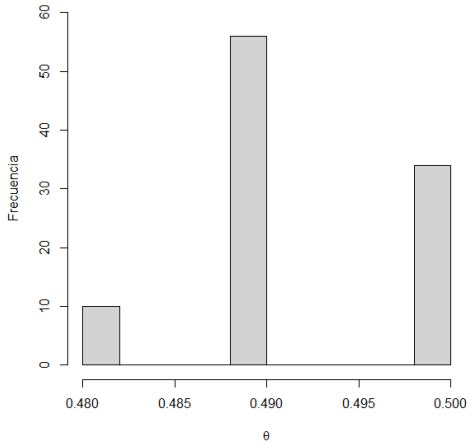
Figura 5.8: La función β está en negro y las estimaciones, en gris.



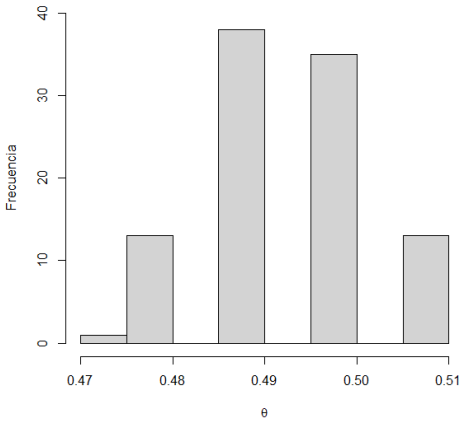
(a) Método A con F_1.



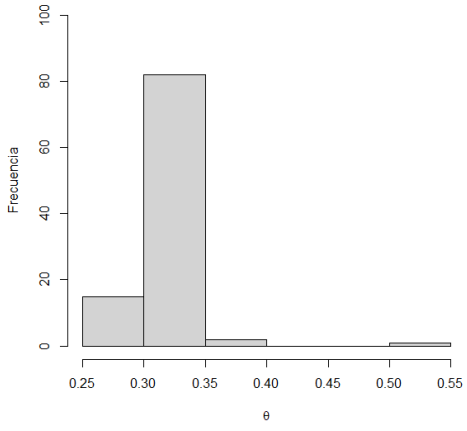
(b) Método A con F_2.



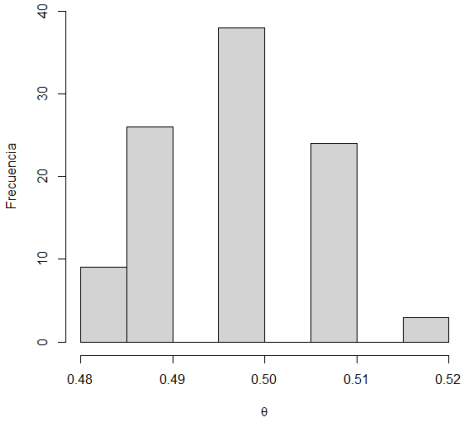
(c) Método A con F_3.



(d) Método B con F_1.



(e) Método B con F_2.



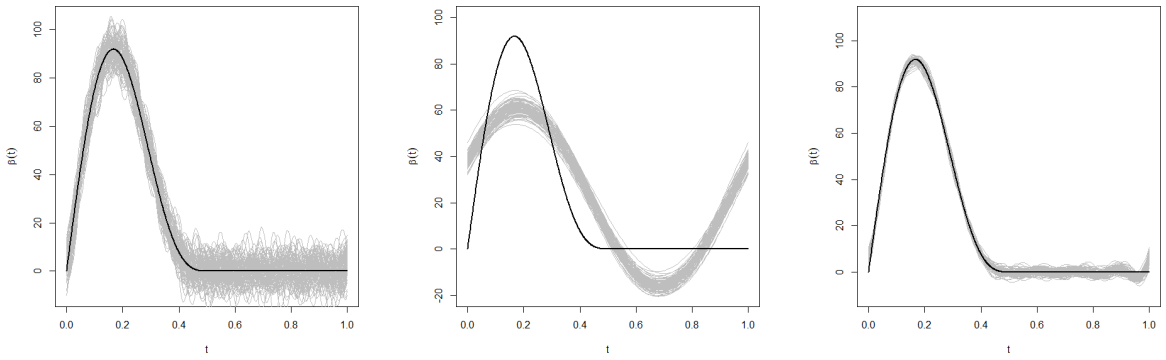
(f) Método B con F_3.

Figura 5.9: Histogramas para $\hat{\theta}$.

5.4. Cuarto modelo

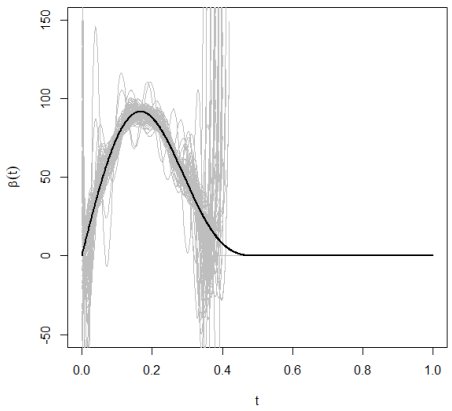
La función β de este modelo es continua con derivadas continuas en $[0, 1]$ hasta orden 3.

Método	Media $\hat{\theta}$	Desv. $\hat{\theta}$	MSE	ISE para β
A (con F_1)	0,385	0,012	0,917	158,441
A (con F_2)	0,377	0,005	3,213	14,742
A (con F_3)	0,398	0,008	0,982	2,662
B (con F_1)	0,414	0,015	0,973	11,226
B (con F_2)	0,376	0,011	55,624	165,320
B (con F_3)	0,420	0,009	1,000	1,428
Sin truncado (con F_1)	-	-	0,737	24,695
Sin truncado (con F_2)	-	-	120,691	323,803
Sin truncado (con F_3)	-	-	0,841	2,529

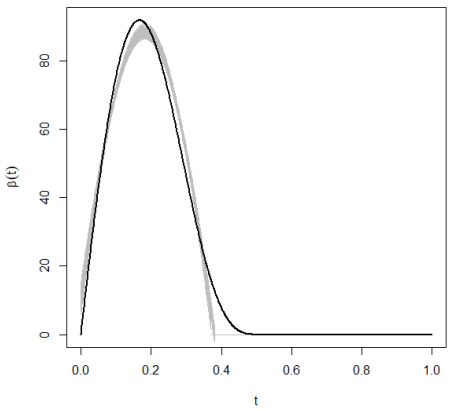


(a) Sin truncado con F_1. (b) Sin truncado con F_2. (c) Sin truncado con F_3.

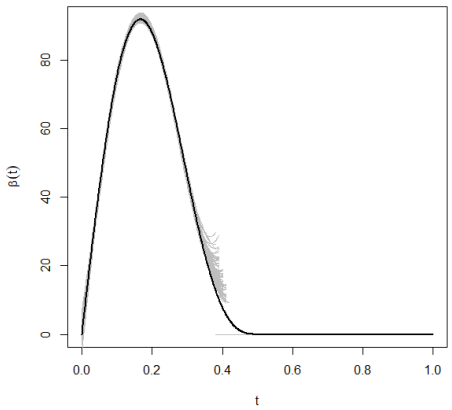
Figura 5.10: La función β está en negro y las estimaciones, en gris.



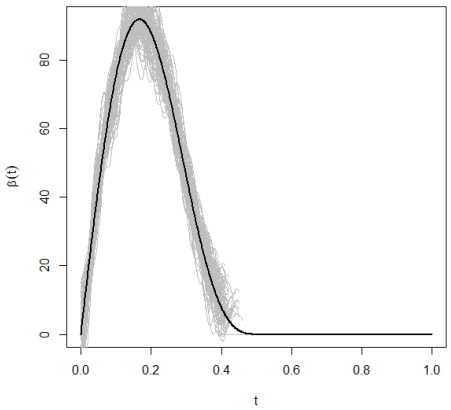
(a) Método A con F_1.



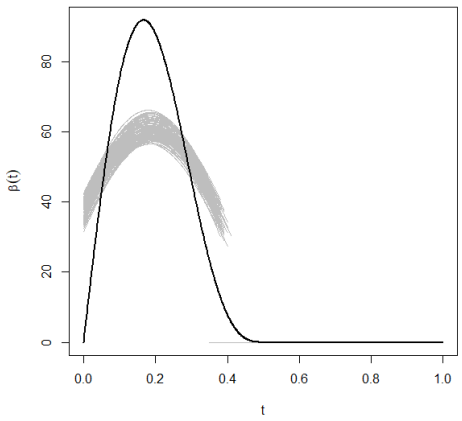
(b) Método A con F_2.



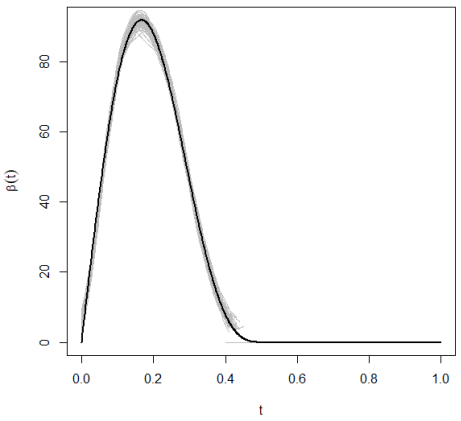
(c) Método A con F_3.



(d) Método B con F_1.

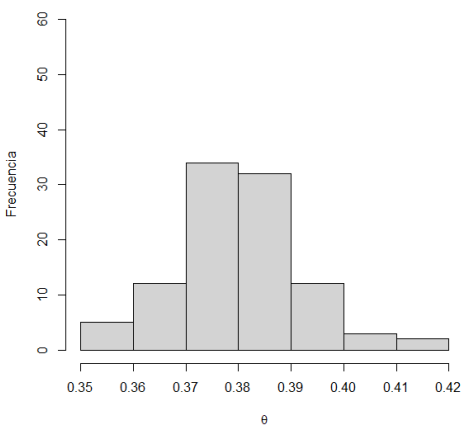


(e) Método B con F_2.

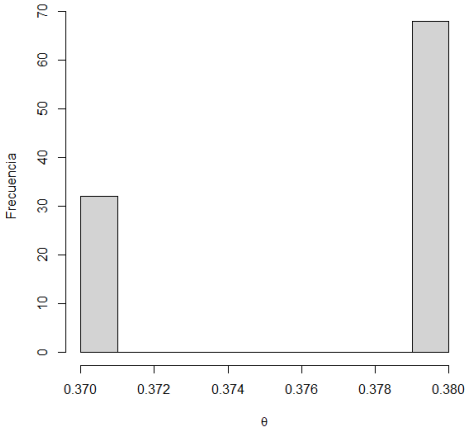


(f) Método B con F_3.

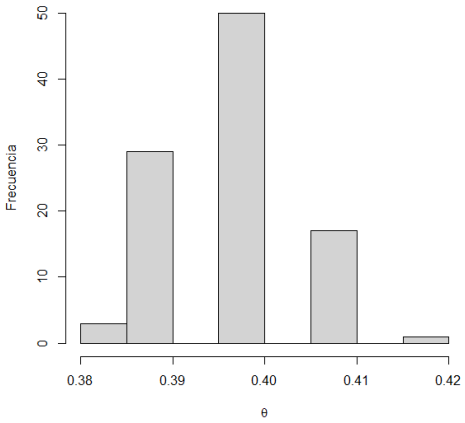
Figura 5.11: La función β está en negro y las estimaciones, en gris.



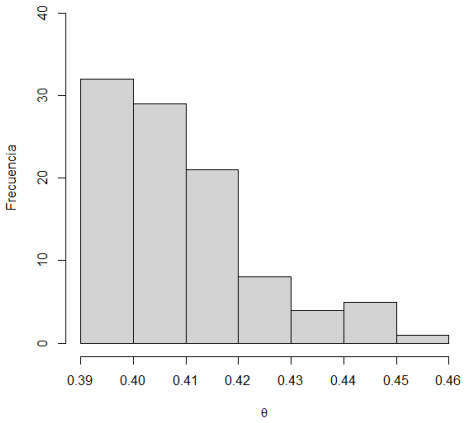
(a) Método A con F_1.



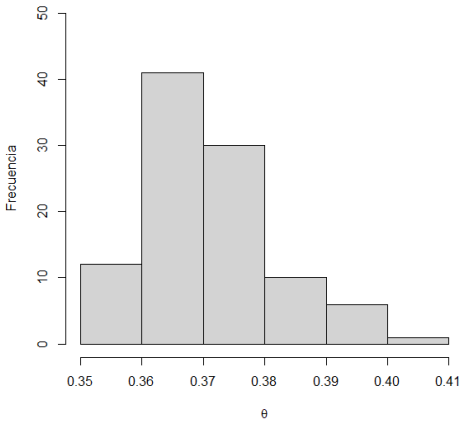
(b) Método A con F_2.



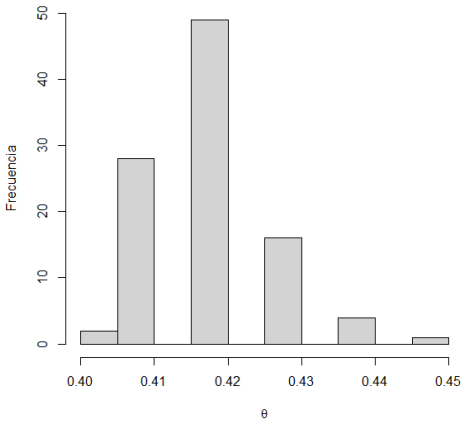
(c) Método A con F_3.



(d) Método B con F_1.



(e) Método B con F_2.



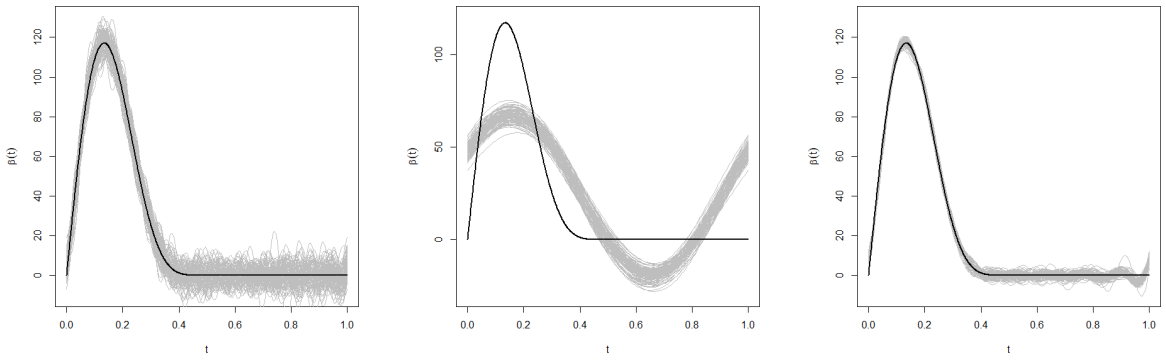
(f) Método B con F_3.

Figura 5.12: Histogramas para $\hat{\theta}$.

5.5. Quinto modelo

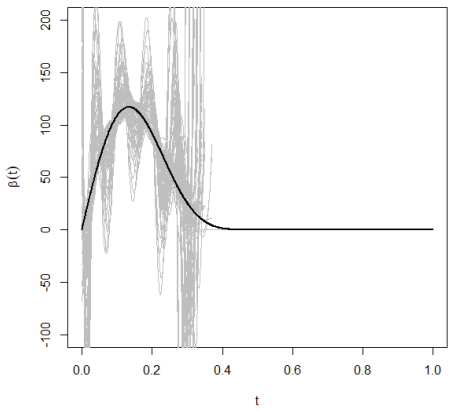
La función β de este modelo es continua con derivadas continuas en $[0, 1]$ hasta orden 5.

Método	Media $\hat{\theta}$	Desv. $\hat{\theta}$	MSE	ISE para β
A (con F_1)	0,329	0,015	$0,908$	1 220,003
A (con F_2)	0,314	0,005	4,555	30,285
A (con F_3)	0,348	0,009	0,979	3,584
B (con F_1)	0,367	0,015	0,983	10,218
B (con F_2)	0,335	0,014	117,090	343,080
B (con F_3)	0,370	0,010	0,980	2,184
Sin truncado (con F_1)	-	-	0,728	24,184
Sin truncado (con F_2)	-	-	234,61	620,527
Sin truncado (con F_3)	-	-	0,856	4,178

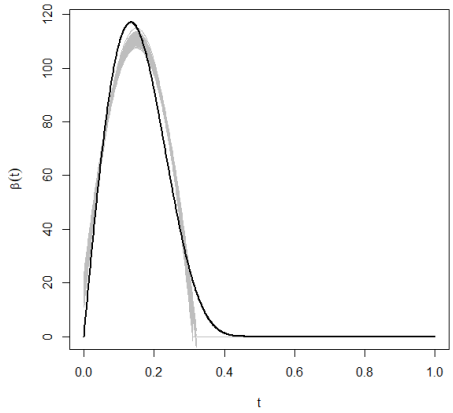


(a) Sin truncado con F_1. (b) Sin truncado con F_2. (c) Sin truncado con F_3.

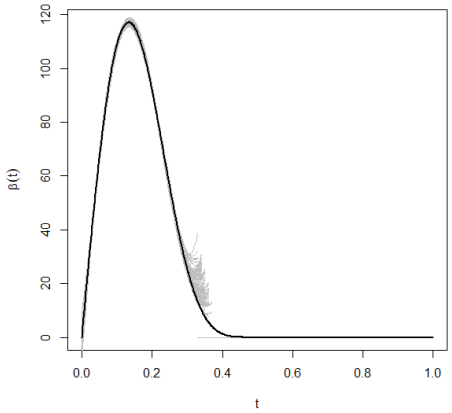
Figura 5.13: La función β está en negro y las estimaciones, en gris.



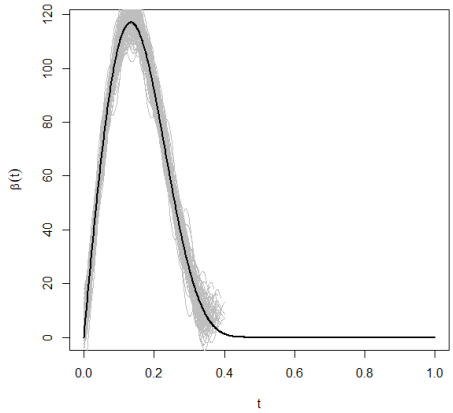
(a) Método A con F_1.



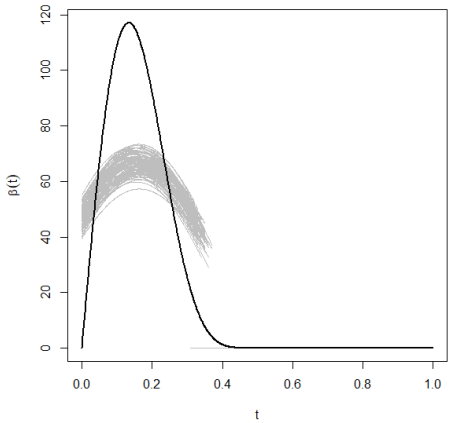
(b) Método A con F_2.



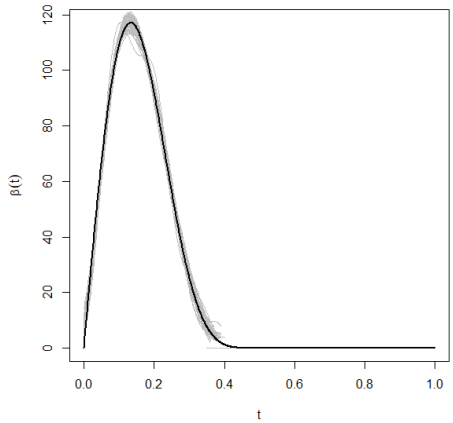
(c) Método A con F_3.



(d) Método B con F_1.

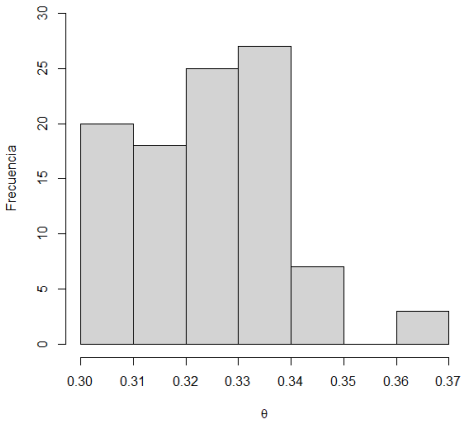


(e) Método B con F_2.

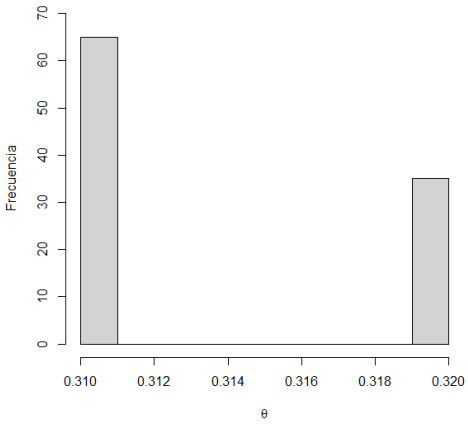


(f) Método B con F_3.

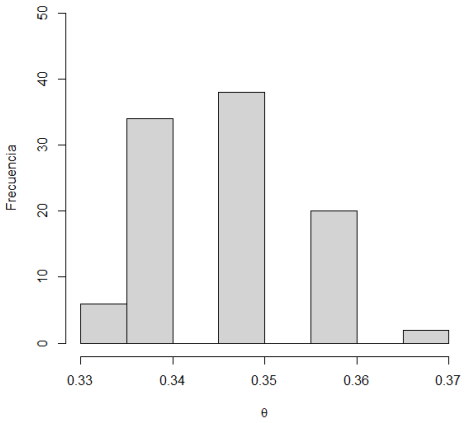
Figura 5.14: La función β está en negro y las estimaciones, en gris.



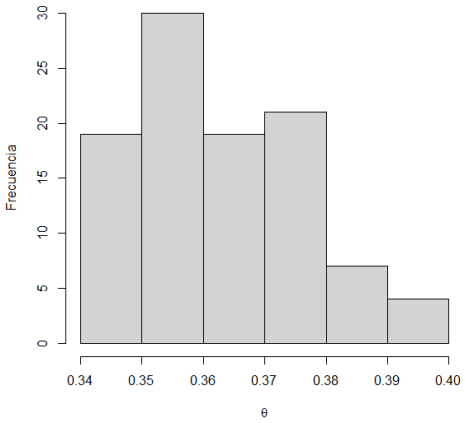
(a) Método A con F_1.



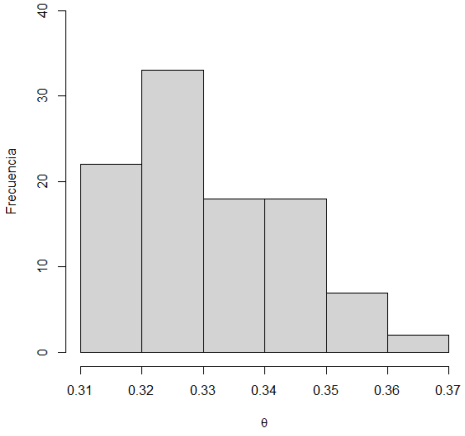
(b) Método A con F_2.



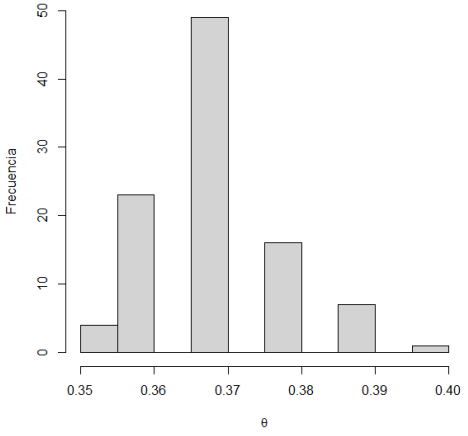
(c) Método A con F_3.



(d) Método B con F_1.



(e) Método B con F_2.



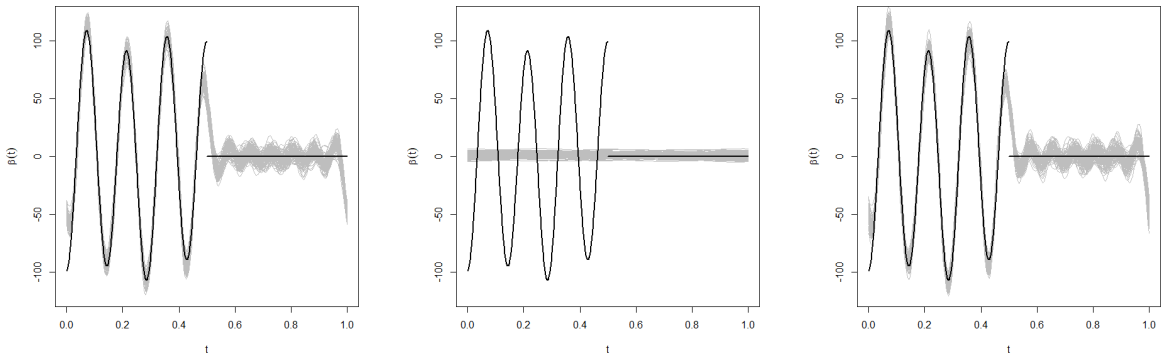
(f) Método B con F_3.

Figura 5.15: Histogramas para $\hat{\theta}$.

5.6. Sexto modelo

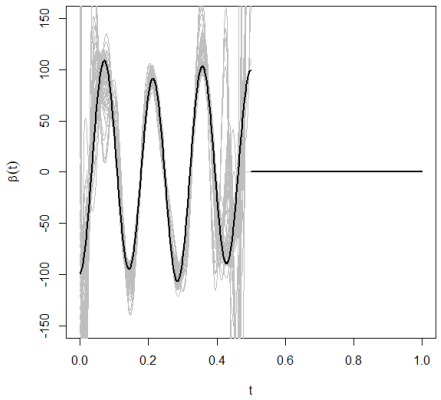
La función β presenta una discontinuidad evitable en θ. Además, presenta ondas de alta frecuencia resultando en derivadas con valores altos.

Método	Media $\hat{\theta}$	Desv. $\hat{\theta}$	MSE	ISE para β
A (con F_1)	0,486	0,006	0,858	3 038,775
A (con F_2)	0,138	0,007	56,244	1 933,448
A (con F_3)	0,492	0,004	0,874	293,697
B (con F_1)	0,511	0,005	1,803	76,288
B (con F_2)	0,181	0,208	88,405	2 467,393
B (con F_3)	0,510	0,003	1,718	75,580
Sin truncado (con F_1)	-	-	0,740	120,476
Sin truncado (con F_2)	-	-	90,845	2 477,142
Sin truncado (con F_3)	-	-	0,713	122,555

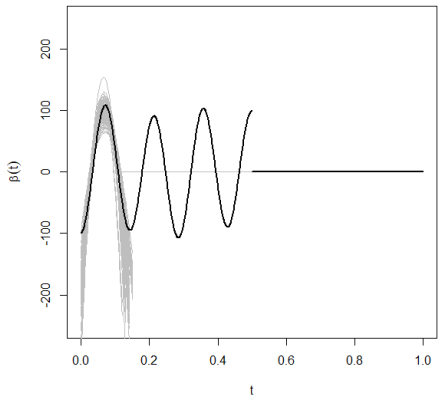


(a) Sin truncado con F_1. (b) Sin truncado con F_2. (c) Sin truncado con F_3.

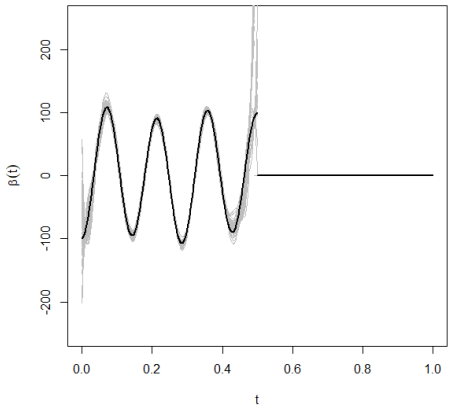
Figura 5.16: La función β está en negro y las estimaciones, en gris.



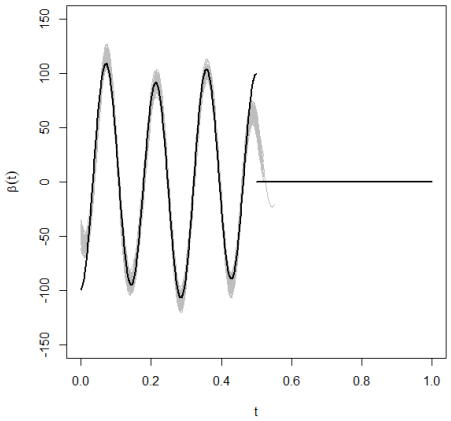
(a) Método A con F_1.



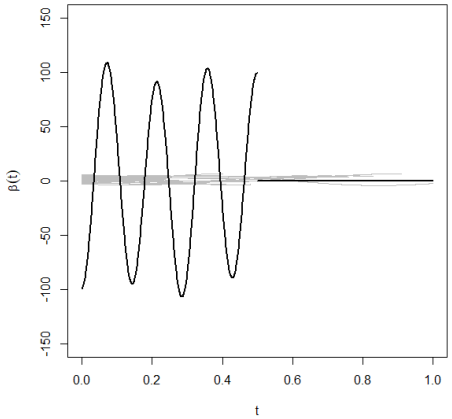
(b) Método A con F_2.



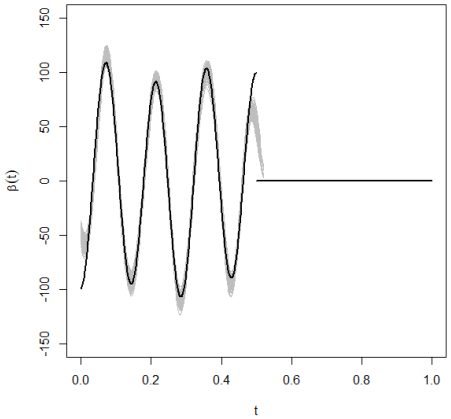
(c) Método A con F_3.



(d) Método B con F_1.

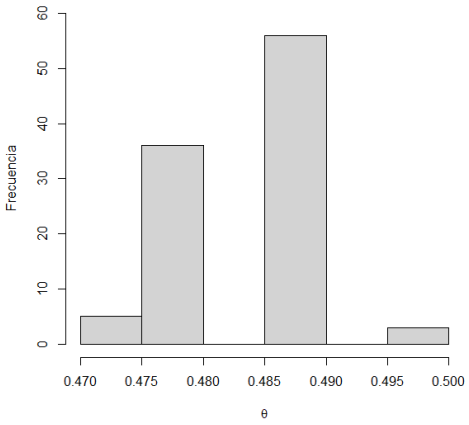


(e) Método B con F_2.

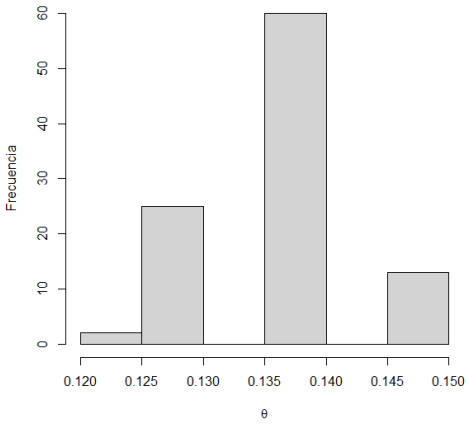


(f) Método B con F_3.

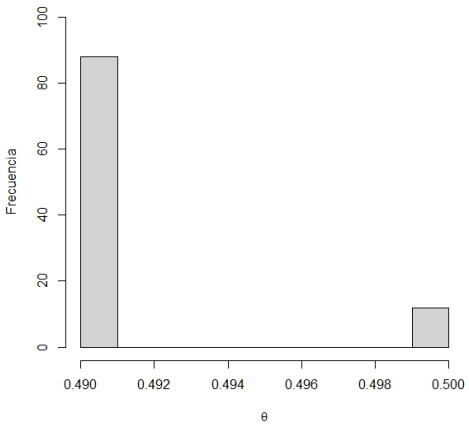
Figura 5.17: La función β está en negro y las estimaciones, en gris.



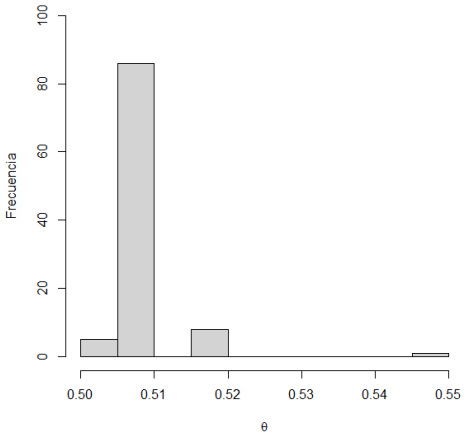
(a) Método A con F_1.



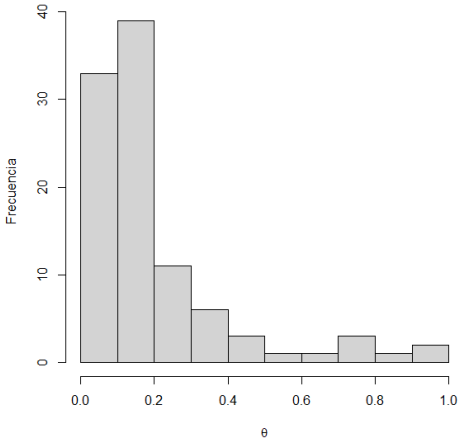
(b) Método A con F_2.



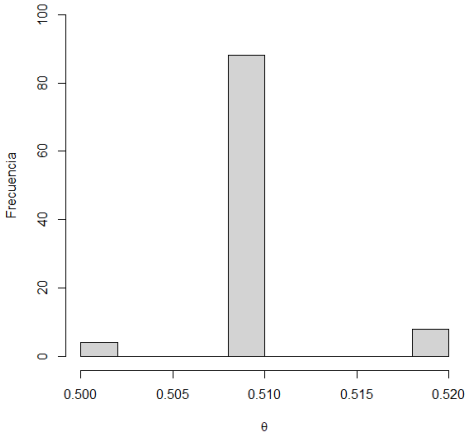
(c) Método A con F_3.



(d) Método B con F_1.



(e) Método B con F_2.



(f) Método B con F_3.

Figura 5.18: Histogramas para $\hat{\theta}$.

Se espera que la elección de F permita algún balance entre el sesgo y la varianza para las estimaciones de β en los diferentes métodos a través de la elección de m. F_1 disminuiría el sesgo sin considerar la varianza. Por otra parte, F_2 tendría en cuenta esto último debido a que involucra a un término adicional $||D\hat{\beta}||_{L^2([0,\hat{\theta}])}$. En consecuencia, la varianza (representada por el ISE) se reduce un poco más a costa de que el sesgo (representado por el MSE) incremente. La última función F_3 pretende solucionar el problema del posible exceso de sesgo que pudiese surgir con F_2. El logaritmo en el término adicional intenta no dar tanto peso al ruido y recuperar la importancia del término referido al sesgo. Esto debería lograr un intermedio entre los resultados de F_1 y F_2.

En los resultados se suele encontrar este patrón esperado en ambos métodos. En el método A suele aparece otro patrón también: la función F_1 induce estimaciones de β mayor varianza mientras que F_3 genera menores y F_2 logra un valor entre ambos. En el método B suele aparece otro patrón en el que F_2 genera $\hat{\beta}$ de mayor sesgo, F_1 de menor y F_3, un valor intermedio.

En ambos métodos y modelos se suele encontrar el siguiente orden respecto al desempeño para estimar θ: F_3 devuelve estimaciones de θ mejores y F_2, más alejadas del valor verdadero. Entre los diferentes modelos, aquellos en donde el verdadero β presenta una discontinuidad en θ generan mejores estimaciones $\hat{\theta}$ que aquellos donde es continua. Un acercamiento lento de β hacia 0 a medida que t tiende a θ en cambio hace que sea difícil detectar θ. Por eso los modelos con β_4 y β_5 no generan resultados tan satisfactorios como en los primeros modelos.

El modelo con β_6 deja en evidencia las complicaciones de la función F_2 para estimarla. Debido a las altas frecuencias de β_6, el término del MSE es despreciable en F_2, en consecuencia, se seleccionan aquellas estimaciones de β con menor monotonía apenas teniendo en cuenta el sesgo.

Respecto a la comparación de los dos métodos, el A parece más satisfactorio para dar estimaciones con menor MSE. Por otra parte, el método B es preferible para obtener estimaciones más cercanas a θ y de menor ISE. Este método entonces puede resultar mejor para predecir respuestas. Esta conclusión no tiene en cuenta al método B mediante el uso de la función F_2 debido a la calidad significativamente menor de las estimaciones.

Bibliografía

[1] Benjamín, M., *Modelo lineal funcional con restricciones de forma*, Tesis de doctorado, 2020.

[2] Bosq, D., *Linear Processes in Function Spaces, Theory and Applications*, Springer, 2000.

[3] Cardot, H., Ferraty, F. y Sarda, P., *Functional linear model*, Statistics Probability Letters 45, 1999.

[4] Craven, P. y Wahba, G., *Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation*, Numerische Mathematik, Springer-Verlag, 1979.

[5] Dalzell, C. y Ramsay, J. *Some Tools for Functional Data Analysis*, Journal of the Royal Statistical Society, 1991.

[6] Debnath, L. y Mikusiński, P., *Introduction to Hilbert Spaces with Applications*, Elsevier, 2005.

[7] Ferraty, F. y Vieu P., *Nonparametric Functional Data Analysis, Theory and Practice*, Springer, 2006.

[8] Gasser, T., Kneip, A., Largo, R. Prader, A. y Ziegler, P., *A method for determining the dynamics and intensity of average growth*, Annals of Human Biology, 1990.

[9] Graves S., Hooker G. y Ramsay J., *Functional Data Analysis with R and MATLAB*, Springer, 2009.

[10] Grenander U. *Stochastic processes and statistical inference*, Arkiv för Matematik 1 No. 17, 1950.

[11] Hall, P. y Hooker G., *Truncated linear models for functional data*, Journal of the Royal Statistical Society, Statistical Methodology Series B, 2015. Suplemento.

[12] Horváth, L. y Kokoszka, P., *Inference for Functional Data with Applications*, Springer, 2012.

[13] Karhunen, K., *Zur Spektraltheorie stochastischer Prozesse*, Annales Academiae Scientiorum Fennicae No. 34, 1946.

[14] Kleffe, J., *Principal components of random variables with values in a seperable hilbert space*, Mathematische Operationsforschung und Statistik 4 No. 34, 1973.

[15] Kokoszka, P. y Reimherr, M., *Introduction to Functional Data Analysis*, CRC Press, 2017.

[16] Malfait, N. y Ramsay, J., *The historical functional linear model*, The Canadian Journal of Statistics Vol. 31 No. 2, 2003.

[17] Ramsay, J. y Silverman, B., *Functional Data Analysis*, Second Edition, Springer, 2005.

[18] Ramsay, J. *When data are functions*, Psychometrika - Vol. 47 No. 4, 1982.