Universidad Nacional Autónoma de México

Facultad de Ciencias

Modelos Aditivos Generalizados

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Actuaria

PRESENTA:

Vanessa Sarai Mora Hernández

TUTORA

Dra. Lizbeth Naranjo Albarrán

Ciudad Universitaria, CD. MX. 2021.

Investigación realizada gracias a la Agencia Estatal de Investigación, España (Proyecto MTM2017-86875-C3-2-R), a la Junta de Extremadura, España (Proyectos IB16054, GR18108 y GR18055), a la Unión Europea (European Regional Development Fondos), y al Programa UNAM-PAPIIT, México (Proyecto IN118720).

Todos los procedimientos realizados en este estudio con participantes humanos se realizaron de acuerdo con los estándares éticos del Comité de Bioética de la Universidad de Extremadura (España) y con la declaración de Helsinki de 1964 y sus posteriores modificaciones.

Los ríos lo saben: no hay prisa.
Vamos a llegar algún día.

Índice general

Introducción

Este trabajo se encuentra dividido en cuatro capítulos; en el primero, titulado "Bases Preliminares", se explican técnicas variadas que nos ayudan a complementar el análisis referente a la modelación, como el tratamiento de valores atípicos, el análisis de componentes principales y la selección de modelos.

En el segundo apartado, "Modelos Lineales Generalizados", se plantea el análisis de modelos lineales, el cual es fundamental —por su diversidad y por su facilidad de análisis— no sólo en el enfoque matemático, sino en diversas áreas como la medicina, la psicología, la química y la economía, entre otras. A saber, los modelos lineales generalizados fueron introducidos en 1972 por Nelder y Wedderburn, de acuerdo con [Yee, 2015], dando paso así a una forma de estimación que considera diversas distribuciones como la Poisson o funciones pertenecientes a la familia exponencial. De este modo, dicha información será introductoria para el capítulo que le sigue.

El tercer capítulo, nombrado "Modelos Aditivos Generalizados" se muestra otro enfoque de modelación para una variable respuesta cuya distribución no es continua; de esta forma, se cumple con la condición de pertenecer a la familia exponencial y se origina la flexiblidad de no explicar únicamente relaciones lineales con ayuda de suavizamientos y de splines.

Finalmente, en el última capítulo, "Parkinson", se expone un ejemplo práctico en el que se busca detectar aquellos pacientes que cumplen con la condición de la enfermedad; para ello se emplean diferentes modelos, tanto lineales como aditivos, que son puestos a competir para buscar el que mejor lo explique.

Capítulo 1

Bases Preliminares

1.1. Reducción de dimensiones

Cuando se cuenta con un gran número de variables, el presentar de manera completa la información con la que estamos trabajando, o en una vista completa, no es la única cuestión a la que nos enfrentamos; a continuación se listan ejemplos de cuando se sugiere aplicar alguna técnica de reducción de dimensión:

- Un base de datos más manejable.
- Optimización de algoritmos debido a reducción de tiempo de procesamiento.
- Reducción de ruido, por ejemplo, reducción de correlación entre variables.
- Ayudar en la interpretabilidad del modelo.

Ya que existen diferentes métodos de reducción de dimensiones, es preciso enlistar algunos con los cuales se puede trabajar en este texto, además de los casos en los que es recomendable aplicarlos:

El primer método es el **Análisis de Componentes Principales (ACP)**, para éste la base de datos es transformado de tal manera que su sistema original de coordenadas $\mathbb{R}^{n_1}$ cambia a $\mathbb{R}^{n_2}$, con $n_1 > n_2$.

El nuevo espacio donde vivirá es seleccionado de acuerdo a la base de datos misma, cada vector será seleccionado de acuerdo a la varianza explicada, siendo así el primer eje la dirección donde se explica la mayor varianza de los datos y así descendentemente; el segundo eje se selecciona de tal manera que sea ortogonal al primer eje, al eje de mayor varianza; así continúa el proceso hasta obtener el mismo número de ejes que se tenía inicialmente. Hasta este punto aún se cuenta con las mismas dimensiones, pues la reducción de éstas se da al seleccionar las componentes nos expliquen el mayor porcentaje de varianza, para que la aportación de ellas sea significativa y la cantidad de información perdida sea pequeña, sin olvidar que cada componente suma $x\,\%$ de la variación total del dataset. El análisis de las componentes se basa principalmente en el cálculo de los eigenvalores y eigenvectores de X, que en este caso es nuestra base de datos acotada a las covariables, cada eigenvector tiene su correspondiente eigenvalor para así obtener el eigenvalor como la dirección del eje, la pendiente; y el eigenvalor como el porcentaje de varianza explicada. El orden de los

eigenvectores será asignado en forma descendente a dicha varianza. Cuando existe correlación en el dataset, aquellas variables que se encuentren correlacionadas positivamente contribuirán fuertemente a la misma componente principal.

Otro método recomendado para la reducción de dimensiones es el **Análisis de Factores (AF)**, para el cual existen variables latentes, que son aquellas que no se observan directamente en los datos pero se infieren a través de ellos. Se considera que la base de datos será una combinación lineal de las variables latentes resultantes y de algún error aleatorio. Además, el número de variables lantentes será menor que el número de variables originales, paso en el cual se concluye la reducción de dimensión.

Un tercer método empleado es el **Análisis de Componentes Independientes (ACI)**, para esta técnica la base de datos se compone por n_1 variables, por lo que es una mezcla de las mismas; a diferencia del análisis de componentes principales, éstas deben de ser independientes y no tienen que estar correlacionadas. El objetivo del ACI es obtener una matriz W, tal que $y = WX$ y además maximiza la independencia de las S componentes obtenidas, tal que las componentes o cualquier transformación de ellas, no esté correlacionada. Se busca además maximizar la no gaussianidad de cualquier combinación lineal de las variables aleatorias.

Para los métodos de ACI y de AF, si se cuentan con menos variables que observaciones, los resultados obtenidos en la reducción de dimensiones será mejor que en las condiciones contrarias.

1.2. Imputación de valores

Los valores faltantes pueden tener un gran impacto en la distribución de los datos, generandose así un sesgo en la información. El principal problema de la imputación de datos es que se cuenta con una visión parcial de la base de datos, por ello existen varias funciones o procedimientos asociados a la manera de controlar el efecto que tienen estos valores en nuestra información generada; en seguida se explica un método que será empleado y se acompaña de un ejemplo.

Varios métodos estadísticos suponen la existencia nula de valores ausentes y son únicamente válidos cuando esta condición se cumple. Amelia es un paquete estadístico de R basado en el llenado de *datasets* incompletos, en el que se analiza la base de datos original y además se evita el sesgo y las aproximaciones incorrectas derivadas de la eliminación de todas las observaciones parcialmente obtenidas del análisis.

El algoritmo mostrado en la figura 1.1 se basa en la imputación múltiple, dicho algoritmo ha mostrado reducir el sesgo y aumentar la eficacia, en comparación con sólo eliminar los registros de la base de datos; sin embargo, métodos como la imputación de la media pueden aumentar el sesgo en la varianza y la covarianza. Algunas de las características de la imputación de datos en este algoritmo son:

- Imputación de valores numéricos.
- Algoritmo basado en Máxima Esperanza Bootstrap.
- Características de validación y de generación de estimaciones más precisas de los valores imputados a utilizar en diferentes aplicaciones, como series de tiempo y estudios transversales.

- Validez de interacción entre variables de la base de datos.
- Después de la aplicación, es posible cualquier aplicación estadística que no considere los valores ausentes. Para combinar los resultados automáticamente en análisis estadísticos se recomienda hacer uso del paquete Zeilig en R.

La imputación múltiple consiste en imputar m valores por cada valor ausente en la base de datos, generando así m *datasets* completos, para cada uno de estos *datasets* los valores observados son los mismos, pero las imputaciones son llenadas con una distribución que refleje la aleatoriedad de éstos.

Cuando se lleva a cabo una imputación múltiple, el primer paso es identificar las variables a incluir en la imputación del modelo, para ello es crucial incorporar tanta información como sea necesaria para el análisis, así como especificaciones que ayuden en la imputación, como la temporalidad de los datos, la interacción entre variables o, en su caso, la especificación de las variables categóricas.

El modelo de imputación supone que los datos completos, considerando incluso los valores ausentes, siguen una distribución normal multivariada con media μ y matriz de covarianza Σ. Para entenderlo mejor, supongamos que M es una matriz con valores $m_{ij} = 1$ si $d_{ij} \in D_{faltantes}$ y 0 en otro caso, una matriz indicadora si el valor es faltante o no. Entonces el siguiente supuesto se basa en

$$\mathbb{P}(M|D) = \mathbb{P}(M|D_{total})$$

De esta forma, los valores ausentes dependen únicamente de los valores con los que se cuentan; cuando esto no se cumple se dice que los datos faltantes son completamente aleatorios, si ocurre así se pueden agregar más variables para cumplir con esta condición.

El algoritmo consiste en utilizar los supuestos anteriores en un Bootstrap de Esperanza Máxima. Para cada paso se aplica el muestreo Bootstrap con el fin de bosquejar el siguiente seleccionado. En cada bosquejo se ejecuta el método de Esperanza Máxima, el cual supone una distribución y unos valores iniciales para la media y la varianza o covarianza, según sea el caso. Con base en ellos se tiene una función esperada de la verosimilitud que será maximizada para obtener los parámetros de esta función. Es importante resaltar que se hace el llenado de cada bosquejo con base en los parámetros obtenidos en el paso anterior y los valores faltantes son calculados utilizando una regresión lineal [Honaker *et al.*, 2012].

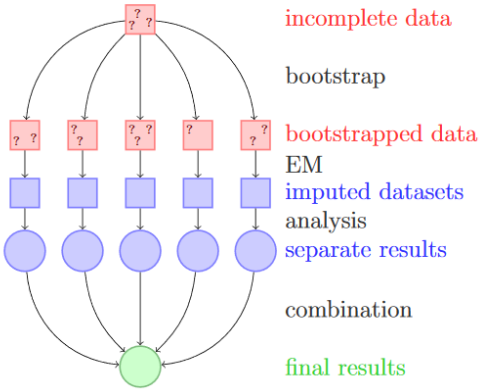


Figura 1.1: Algoritmo de Amelia.

1.3. Selección del modelo ganador

Existen diferentes maneras de poner a competir la modelación con variaciones en el número de variables, la forma de las variables o incluso el tipo de algoritmo utilizado, para ello se presentan dos tipos: basados en el error o basados en la verosimilitud.

1.3.1. Medidas basadas en el error

- **Validación cruzada.** La validación cruzada es un método que evalúa la tasa de error de un modelo. Para esta medida se divide aleatoriamente la base de datos original en un subconjunto de entrenamiento y de prueba.

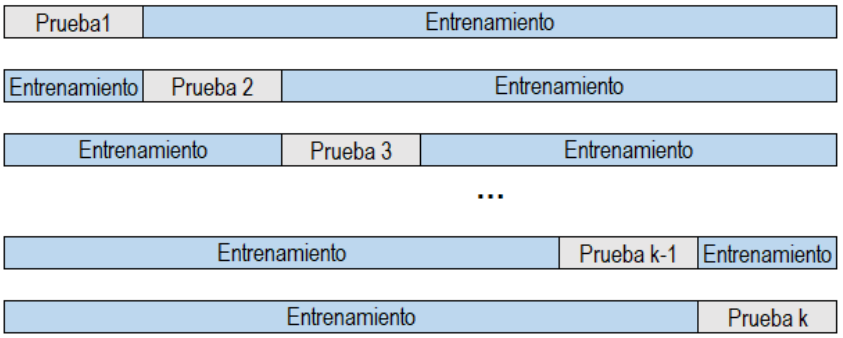


Figura 1.2: Representación de la selección de las particiones aleatorias.

La base de datos de entrenamiento es aquel con el cual se calibrará el modelo que hará futuras predicciones. A su vez, la base de datos de prueba será evaluado con el modelo generado anteriormente, asimismo, las estimaciones resultantes serán comparadas con los valores originales, es aquí donde se calcula el error. Esto se repite k veces, de tal manera que el score final que tendrá el modelo será dado por el promedio:

$$\epsilon = \sum_{i=1}^{n} \frac{e_i}{k}$$

Para las siguientes medidas es necesario calcular la matriz de confusión, la cual muestra una tabla de contingencia donde se compara la variable respuesta del modelo o predicción calculada contra los datos originales o los datos observados. La matriz es dada como se muestra en la tabla 1.1.

	Predicción 0	Predicción 1
Observado 0	Verdaderos Negativos (VN)	Falsos Positivos (FP)
Observado 1	Falsos Negativos (FN)	Verdaderos Positivos (VP)

Tabla 1.1: Matriz de confusión.

De la tabla 1.1 contamos con cuatro secciones:

- **Verdaderos Negativos.** Son aquellos observados que no contaban con la condición y el modelo los clasificó correctamente.
- **Verdaderos Positivos.** Son el conjunto de predicciones que cumplen con la condición a modelar y en la realidad se refleja de igual manera.
- **Falsos Negativos.** Se refire a aquellos que el modelo no logró detectar con la condición dada, pero que realmente la tenían. Por lo general, éstos afectan la efectividad del algoritmo empleado.
- **Falsos Positivos.** Son los que el modelo clasificó como aquellos que tenían la condición a calcular y realmente no contaban con ella.
- **Sensibilidad.** La sensibilidad se calcula de la siguiente manera:

$$Sensibilidad = \frac{VP}{VP + FN}$$

  Entre más cercana sea a 1, implica que el modelo clasificó correctamente el porcentaje de aquellos que contaban con la condición al contar con un error muy pequeño en aquellos que no contaban con la condición y el modelo los clasificó como si contaran con ella, es decir, los falsos negativos.

- **Especificidad.** La especificidad calcula la proporción de aquellos que no presentaban la condición y el modelo los clasificó correctamente, contra el total de los no que cuentan con la condición e incluyendo aquellos mal clasificados, es decir, los falsos positivos.

$$Especificidad = \frac{VN}{VN + FP}$$

Por cada punto de corte, la sensibilidad y la especificidad cambian; cuando nuestro objetivo es buscar el punto de corte óptimo, debemos encontrar aquel punto que maximice ambas medidas. La curva ROC, (Receiver Operating Characteristic, por sus siglas en inglés) grafica ambas medidas, Sensibilidad vs $1 - Especificidad$ para todos los puntos de corte posible, estos puntos de corte van asociados a la probabilidad sobre la cual la transformación se verá reflejada como 0 o 1. El área bajo la curva nos proporciona una medida sobre la forma en la predice el modelo. La forma de medir la eficiencia del modelo a través de la curva ROC se puede resumir en:

- Si $ROC \leq 0.5$ entonces el modelo no ayuda a discriminar.
- Si $0.6 \leq ROC \leq 0.8$ entonces el modelo discrimina de manera adecuada.
- Si $0.8 < ROC \leq 0.9$ entonces el modelo discrimina de manera excelente.

1.3.2. Medidas basadas en la verosimilitud

Cuando estimamos modelos, es posible aumentar la verosimilitud agregando parámetros, pero al hacerlo se puede caer en un sobreajuste del modelo, para ello se calculan las siguientes medidas:

- **Criterio de información de Akaike.** Se encuentra definido por la siguiente fórmula:

$$AIC = -2\log(L) + 2k$$

  Una función basada en el logaritmo de la verosimilitud y una penalización basada en el número de parámetros del modelo, determinado por k. Cuando el criterio de Akaike es el menor, podemos decir entonces que encontramos un modelo ganador, debido a que entre menor sea el valor obtenido, mejor es el ajuste del modelo, ya que para este criterio no sólo se toma la información de los datos, sino también se considera si el modelo es parsimonioso o no.

- **Criterio de información de Bayes.** También conocido como el criterio de información de Schwarz, éste refiere a una función creciente, la cual depende de la varianza no explicada entre la variable dependiente y el número de variables explicativas. Por lo que, cuando el valor del criterio de información de Bayes es bajo, puede implicar que el modelo cuenta con pocas variables, que el modelo no está sobreajustado o ambas situaciones:

$$BIC = -2\log L + k\log n$$

  A diferencia del Criterio de Akaike, éste penaliza en mayor proporción el contar con una gran cantidad de variables dependientes; de igual forma, es importante resaltar que para comparar modelos usando este criterio es necesario que dichos modelos tengan la misma cantidad de observaciones.

Capítulo 2

Modelos Lineales Generalizados

2.1. Modelos Lineales Generalizados

Usualmente se asocia el término de modelos lineales con una regresión lineal, en este capítulo se mostrará la generalización de éstos y su aplicación. Al hablar de modelos lineales generalizados necesitamos preguntar varios conceptos o plantear diversas preguntas: ¿Qué son?, ¿Cómo me pueden ayudar?, ¿Son aplicables a todos los datos?, ¿Cómo saber que estoy generando un buen modelo o estimación con base en ellos?

En estadística se utilizan modelos lineales para explicar la relación entre covariables y una variable respuesta, pueden ser aplicables a pares de datos de la forma (y, X) debido a que nos permiten estudiar patrones, variaciones y relaciones lineales entre las covariables y la variable respuesta, la cual sigue una distribución aleatoria.

Usualmente, esto se representa mediante las ecuaciones

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

donde para $i = 1, \ldots, n$, con n tamaño de muestra, y_i representa la variable respuesta o variable objetivo, la cual intentamos modelar mediante las covariables o variables explicativas, $x_i = (x_{i1}, \ldots, x_{ip})'$, la forma en que se relaciona cada una de las covariables con y_i está dada por $\beta = (\beta_1, \ldots, \beta_p)'$, y ε_i son errores aleatorios se basa en el supuesto de ser identicamente distribuidos bajo una función F, con media cero y varianza constante denotada por σ^2 Los parámetros estimados serán $\hat{\beta}_1, \ldots, \hat{\beta}_p$ y el modelo estimado estará dado por

$$\hat{y}_i = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$$

Si bien no se llega a medir completamente la relación entre y_i y $\hat{y}_i$ nos podemos aproximar linealmente y conocer qué tan correcta es nuestra estimación mediante el cálculo de los errores:

$$y_i = \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_p x_{ip} + \hat{\varepsilon}_i$$

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Cuando un modelo lineal no ajusta de manera adecuada o únicamente no cumple el supuesto de normalidad, se pueden encontrar distintas maneras de solventar el problema; la manera más común es mediante el uso de transformaciones para aproximar a la distribución

Normal. Los modelos lineales generalizados incluyen casos especiales como una regresión lineal o un análisis de varianza, pero van más allá al estimar modelos para la familia exponencial incluyendo así variables respuestas no continuas y además generando, por ejemplo, un modelo logit, modelos para la distribución Poisson, entre otros.

Para la parte aleatoria Y, es independiente y con varianza constante en los errores, además de considerar a $f(y;\theta)$ como la función de densidad de acuerdo con los parámetros que se están estimando. El proceso de estimación del modelo es definido por alguna medida de bondad de ajuste, la cual nos ayudará a la interpretación de la distancia que separa nuestra estimación del valor real $(y_i - \hat{y}_i)$; la métrica de la distancia va en función de los errores $\hat{\varepsilon}_i$:

$$S(y_i, \hat{y}_i) = \sum (y_i - \hat{y}_i)^2$$

Se busca mediante una estimación por mínimos cuadrados aquellos parámetros que minimicen la distancia de los errores, además se minimiza la ecuación anterior para obtener el mejor estimador lineal que no esté sesgado; a su vez, cuando se agrega el supuesto $\varepsilon_i \sim \mathrm{N}(0,\sigma^2)$ coincide con la máxima verosimilitud, de esta manera se obtiene el modelo ganador: aquel que minimice la ecuación anterior; no obstante, se debe de tener cuidado, puesto que no funciona bien en todos los casos debido a la sensibilidad en:

- Outliers.
- Colinealidad.
- Cuando $p > n$, es decir, la dimensión de $X = (x_1, \cdots, x_p)$ es mayor a la de y_i, lo cual implica que existen más variables explicativas que observaciones a modelar.

Por otro lado, un Modelo Lineal Generalizado se compone de tres partes:

- Componente aleatoria: Variable aleatoria Y, cuya función de distribución o de probabilidad es denotada por $f(y_i)$.
- Predictor lineal: $\eta = X\beta$, compuesto por el vector de parámetros $\beta = (\beta_1, ..., \beta_p)$ y la matriz de covariables X.
- Función liga (función liga o de enlace): $g(E[y]) = g(\mu) = X\beta$, la cual asocia cada componente de $E[y]$ con el predictor lineal.

Los modelos lineales generalizados (GLM) se encuentran restringidos a modelos cuya variable objetivo (y) pertenece a la familia exponencial; una regresión o un modelo lineal es un caso particular, donde la variable respuesta sigue una distribución Normal. Las covariables pueden ser de dos tipos:

- Cuantitativas: son las variables más comunes para el desarrollo del modelo, pueden ser numéricas continuas o de intervalo.
- Cualitativas: son variables categóricas, ordinales, indicadoras. En estos casos es común generar variables dummies (variables indicadoras que toman valores 1 o 0), una por cada categoría de la variable, menos una que será la referencia.

Sea y, con media μ, consideramos el GLM $\eta = X\beta$, con función liga $g(\mu) = \eta$ donde y, μ y η pertenecen a $\mathbb{R}^n$. Sea $\mathcal{S}$ un espacio vectorial tal que:

$$\mathcal{S} = \{\eta \mid Existe\, \beta\, que\, cumple\, \eta = X\beta\}$$

Se observa que la dimensión de $\mathcal{S}$ es definida por el rango de X, el rango también puede ser interpretado como el número de covariables independientes con las que cuenta el modelo; cuando el rango de X es diferente de su número de columnas, implica la existencia de colinealidad en los datos, es decir, alguna columna x_k es combinación lineal de una o más columnas de X, al ocurrir esto S se vuelve un espacio vectorial nulo y no es posible trabajar con X para el modelo, por lo que se debe de tener cuidado con esta condición.

2.1.1. Función liga

La función liga, como se presentó anteriormente, conecta la componente aleatoria con el predictor lineal, además es la función que diferencia un modelo lineal generalizado de un modelo de regresión común. Sea $\mu_i = E[y_i]$, con $i = 1, \cdots, n$.

- La función liga conecta η_i a μ_i de la forma $\eta_i = g(\mu_i)$, donde g(.) cumple con ser monótona y derivable.
- La función liga que transforma μ_i al parámetro natural es llamada **función liga canónica**.
- Cuando $g(\mu_i) = \mu_i$, es llamada **función liga identidad**, a su vez, en el momento que un modelo lineal generalizado utiliza esta función, se sigue:

$$g(\mu_i) = \mu_i = \sum_{j=1}^{p} \beta_j x_{ij}$$

  si además se cumple la homocedasticidad, media constante y el error sigue una distribución normal, la función anterior representa un modelo lineal.

2.1.2. Familia Exponencial de un parámetro

Sea f una función de densidad, decimos que $f(y_i; \theta, \phi)$ pertenece a la familia exponencial de un parámetro si tiene la forma:

$$f(y_i; \theta_i, \phi) = e^{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)} \ldots (2.1)$$

donde:

- θ_i es el parámetro natural o canónico de la familia.
- ϕ es el parámetro de escala de la función.

Usualmente $a(\phi_i) = 1$ o conocida y $c(y_i, \phi) = c(y_i)$, cuando no se cumple este supuesto la ecuación pertenece a una distribución de dos parámetros, como una Normal(μ, σ^2) o una Gamma(μ, υ).

Ahora bien, deseamos conocer aquellos sitios donde la función cuenta con puntos criticos, por lo que procedemos a calcular la log-verosimilitud de la función (2.2), teniendo así:

$$L = \sum \log(f(y_i; \theta_i, \phi)) = \sum \left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right) \ldots (2.2)$$

Además la función cumple:

$$\mathbb{E}\left[\frac{d\mathcal{L}}{d\theta}\right] = 0$$

y

$$\mathbb{E}\left[\frac{d^2\mathcal{L}}{d\theta^2}\right] + \mathbb{E}^2\left[\frac{d\mathcal{L}}{d\theta}\right] = 0$$

Para la representación de la E[y_i] y la Var(y_i), se utiliza la forma de la ecuación (2.2) de la familia exponencial. Sea $\mathcal{L}_i = log(f(y_i; \theta_i, \phi))$ y sea $L = \Sigma\mathcal{L}_i$. Se sigue entonces

$$\mathcal{L}_i = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \tag{2.3}$$

$$\frac{\partial\mathcal{L}_i}{\partial\theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} \tag{2.4}$$

$$\frac{\partial^2\mathcal{L}_i}{\partial\theta_i^2} = \frac{b''(\theta_i)}{a(\phi)} \tag{2.5}$$

Si se sigue el mismo proceso que para la verosimilitud, por tanto:

$$\mathbb{E}\left[\frac{\partial\mathcal{L}_i}{\partial\theta_i}\right] = 0 \tag{2.6}$$

$$-\mathbb{E}\left[\frac{\partial^2\mathcal{L}_i}{\partial\theta_i^2}\right] = \mathbb{E}^2\left[\frac{\partial\mathcal{L}_i}{\partial\theta_i}\right] \tag{2.7}$$

De (2.6) se concluye

$$\mu_i = \mathbb{E}[y_i] = b'(\theta)$$

y de (2.5) y (2.7) se sigue

$$\frac{b''(\theta_i)}{a(\phi)} = \mathbb{E}^2\left[\frac{y_i - b''(\theta_i)}{a(\phi)}\right] = \frac{var(y_i)}{a(\phi_i)}$$

$$\Rightarrow var(y_i) = b''(\theta_i)$$

De tal manera que, al contar con el parámetro natural, es posible obtener la esperanza y la varianza de la distribución y vicerversa.

Para ver la aplicación de lo anterior, se resume en la siguiente tabla la descomposición de la función de densidad de una variable aleatoria cuya distribución sea Poisson(μ), Binomial(n, π) o Normal(μ, σ^2).

Distribución:	Poisson(μ)	Binomial(n, π)	Normal(μ, σ^2)
$f(y_i)$	$\frac{e^{-\mu}\mu^{y_i}}{y_i!}$	$\binom{n}{ny_i}\pi^{ny_i}(1-\pi)^{n-ny_i}$	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp -\frac{(y_i-\mu)^2}{2\sigma^2}$
θ	$\log(\mu)$	$\log(\frac{\pi}{1-\pi})$	μ
$a(\phi)$	1	$\frac{1}{n}$	σ^2
$b(\theta)$	$e(\theta)$	$\log(1+\exp(\theta))$	$\frac{1}{2}\theta^2$
$c(y,\phi)$	$-\log(y_i!)$	$\log\binom{n}{ny_i}$	$-\frac{\log(2\pi\sigma^2)}{2}-\frac{y_i^2}{2\sigma^2}$

Tabla 2.1: Resumen de la descomposición en componentes de la familia exponencial.

Para el caso de la distribución Binomial se emplea una parametrización diferente a la usual, donde y_i representa la proporción de éxitos, lo que implica que la $\mathbb{E}[y_i] = \pi$.

De acuerdo con lo tratado en esta sección y con base en la información de los modelos lineales generalizados, se tienen diversas opciones para la función liga; una de las propuestas es utilizar la función liga canónica dada por la transformación a la media por el parámetro canónico correspondiente, es decir:

$$\eta = g(\mu) = \theta$$

2.2. Modelos de regresión binaria

Un modelo de regresión logística es un modelo lineal generalizado donde la variable respuesta es binaria, es decir: éxito/fracaso, 0/1, la cual puede describirse como un evento Bernoulli. Si el número de ensayos es mayor a 1, entonces y sigue una distribución Binomial(n,p).

Nuestro modelo se compone entonces de:

- $x_i = [1, x_{i1}, ..., x_{ip}]$, con $i = 1, \cdots, n$
- $\beta = [\beta_0, \beta_1, ..., \beta_p]$
- $y_i = \begin{cases} 0 \\ 1 \end{cases}$

Un evento Bernoulli para la observación i tiene:

- $\mathbb{P}(y_i = 1) = \pi_i$
- $\mathbb{P}(y_i = 0) = 1 - \pi_i$

Existe una dependecia de π_i con x_i, la cual ocurre a través de la combinación lineal:

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$$

Entonces la función liga es expresada de la siguiente manera:

$$g(\pi_i) = \eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$$

Para que sea consistente con la definición de probabilidad se utiliza la función liga la cual mapea el intervalo [0,1] que es donde corre π_i al intervalo $(-\infty, \infty)$. La función liga nos dará el tipo de modelo.

Las funciones liga más utilizadas son:

- Para la regresión logística:
$$g(\pi) = log(\frac{\pi}{1-\pi})$$
- Para la regresión probit:
$$g(\pi) = \Phi^{-1}(\pi)$$
- Para la regresión complementaria log-log o cloglog:
$$g(\pi) = log(-log((1-\pi))$$

Ahora bien, definamos ${y_i}^* = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$, donde ε_i sigue una distribución F, con media cero. Entonces, existe un límite o barrera τ tal que $y_i = 0$ si ${y_i}^* \leq \tau$ y $y_i = 1$ si ${y_i}^* > \tau$. Se sigue entonces

$$\pi_i = P(y_i = 1) = F\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)$$

$$\Rightarrow F^{-1}\left(P(y_i = 1)\right) = \sum_{j=1}^{p} \beta_j x_{ij}$$

Por lo general $\tau = 0$ y F corresponde a una función de probabilidad acumudada que además es simétrica al rededor del cero, por lo que es correcto $F(z) = 1 - F(z)$. Y además:

$$P(y_i = 1) = P({y_i}^* > \tau) = P\left(\sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i > \tau\right)$$

$$= 1 - P\left(\varepsilon_i \leq \tau - \sum_{j=1}^{p} \beta_j x_{ij}\right)$$

$$= 1 - F\left(\tau - \sum_{j=1}^{p} \beta_j x_{ij}\right)$$

En la siguiente imagen se observa un ejemplo sobre cómo representar gráficamente este límite τ:

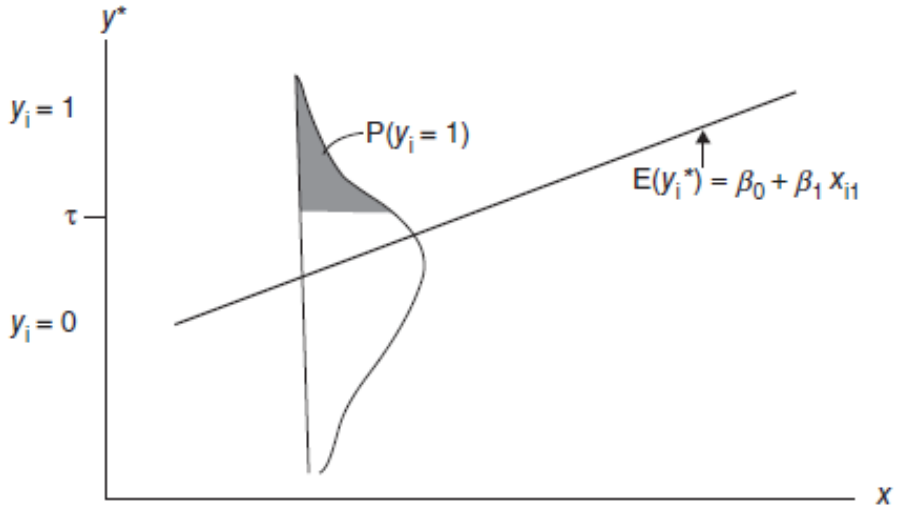


Figura 2.1: Distribución Normal para un modelo probit.

Siguiendo lo anterior, cuando F corresponde a una distribución normal estándar, entonces F^{-1} es referida a la función inversa de la distribución Normal o bien la función probit y si además se toma en cuenta:

$$\pi_i = P(y_i = 1) = F\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)$$

$$\Rightarrow F^{-1}[P(y_i = 1)] = \sum_{j=1}^{p} \beta_j x_{ij}$$

Al cumplirse esto, se conoce como un modelo de regresión probit y F^{-1} funje como la función liga del modelo.

En el caso de la regresión logística, F corresponde a una distribución logistic con media cero, entonces F^{-1}, al estar asociada a la función link, es la función logit. Por lo que:

$$\pi_i = \frac{\exp(\sum_{j=1}^{p} \beta_j x_{ij})}{1 - \exp(\sum_{j=1}^{p} \beta_j x_{ij})}$$

$$\Leftrightarrow logit(\pi_i) = log\left(\frac{\pi_i}{1 - \pi_i}\right) = \sum_{j=1}^{p} \beta_j x_{ij}$$

La forma en la que influyen las β_j representa la tasa de cambio de la función logit por unidad de cambio de cada variable dependiente x_j, para la interpretación contamos con dos casos generales:

- Para variables cuantitativas, la magnitud de β_j la pendiente de la curva es aquella que describe la tasa de cambio de la probabilidad π_i, dada por:

$$\frac{\partial \pi_i}{\partial x_{ij}} = \beta_j \frac{\exp(\sum_j \beta_j x_{ij})}{[1 + \exp(\exp(\sum_j \beta_j x_{ij}))]^2} = \beta_j \pi_j (1 - \pi_j)$$

- Para variables cualitativas cada coeficiente β_j del modelo $logit(\pi_i)$ se puede transcribir como una tabla de contingencia, donde la probabilidad está definida por $\frac{\pi_i}{1-\pi_i}$. Esta última aumenta en múltiplos de $\exp(\beta_j)$ por unidad en x_j.

Si no se cumplen los supuestos al realizar esta modelación se puede caer en alguno de los siguientes errores:

- **Coeficientes sesgados:** dado por sobreestimar o por subestimar los coeficientes del modelo.
- **Estimadores Ineficientes:** se da cuando el valor del error estándar para los coeficientes es alto, teniendo así que el parámetro asociado es cero.
- **Inferencia estadística no válida:** surge cuando se tienen valores atípicos o valores extremos en las variables independientes.

Anteriormente describimos la regresión logística y probit que, además de ser las más utilizadas, tienen un ajuste similar, sin embargo no son las únicas funciones liga. Es posible generar la transformación utlizando como liga la función cloglog o log-log complentaria, esta última se vincula de la siguiente manera:

$$\eta = log[-log(1 - \pi_i)]$$

Y así se sigue que al utilizar esta función liga, la probabilidad ajustada se traduce como:

$$\pi_i = \exp\left(-\exp[\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n]\right)$$

Finalmente, en este capítulo se muestran funciones liga diferentes a utilizar, al ser las tres principalmente utlizadas o programadas en algún software estadístico, cada una con sus pros y contras. Sin embargo no se debe limitar sólo a estas funciones como funciones liga.

Capítulo 3

Modelos Aditivos Generalizados

3.1. Introducción

A pesar de que en el capítulo anterior se presentó una solución a problemas donde la variable respuesta no es continua, se mantiene, sin embargo, el supuesto de que la influencia de las covariables en la variable respuesta es lineal, pero ¿qué pasa si no es necesariamente así?

Para ello es posible ocupar suavizamientos, éstos tienen muchos propósitos generales como visualización o exploración de los datos, predicción o estimación de cómo puede ser el crecimiento de la curva. Por lo tanto, un suavizamiento es una herramienta para resumir la tendencia de la variable respuesta, la cual depende de una o más covariables, asimismo, produce una estimación con menor varianza que la Y original.

Los suavizamientos son utilizados usualmente con dos propósitos: para efectuar un análisis descriptivo mediante un scatterplot, debido a que es más fácil visualizar y detectar el comportamiento que siguen los datos por analizar, delimitado a una visión bilateral; y como estimador en modelos no paramétricos o modelos donde se desea estimar la dependencia de la media de Y y de algunas covariables, que es el caso en el que nos encontramos al poder vincularlo con η.

Un ejemplo de suavizamiento puede ser simplemente el promedio, el cual no necesariamente es pensado como un suavizamiento, comúnmente estos promedios son realizados en vecindades alrededor del valor a estimar, por ello nos enfrentamos a varios problemas: ¿cómo calcular la respuesta promedio en cada vecindad? ¿Cuál es el tamaño óptimo de dicha vecindad?

El parámetro del suavizamiento es también conocido como el tamaño de la vecindad a calcular, si éste es grande se generan estimaciones con varianza pequeña, pero seguramente muy sesgadas, a tal grado que podría ser más conveniente no utilizar un suavizamiento, por lo cual la segunda pregunta se resume en: ¿cuál es la forma óptima de calcular el parámetro del suavizamiento?

Existen cuatro tipos de suavizamientos que se muestran a continuación:

- Suavizamientos para regresiones o series.
- Splines de suavizado.
- Regresiones locales.

- Suavizamiento para el vecino más cercano.

El primer tipo, el **suavizamiento para regresiones** o para series de tiempo, es una técnica basada en el ajuste de polinomios donde cada x_k da mayor flexibilidad comparado con un regresor lineal normal ($\beta_k x_k$), asimismo este método es fácil de interpretar cuando el polinomio es de grado pequeño. De acuerdo con el teorema de aproximación de Weierstrass:

Teorema 1 *Sea $f : [0, 1] \to \mathbb{R}$ una función continua, existe una sucesión de polinomios (p_n) que converge uniformemente a f en $[a, b]$.*

Se puede caer en ajustar un polinomio de grado muy alto, no obstante, un punto a considerar es que al aumentar el grado del polinomio también se pierde la parsimonia del modelo, inclusive este método es sensible a valores atípicos, por lo que podría darse un sobreajuste del mismo.

En el caso del segundo tipo, los **splines de suavizado** controlan la flexibilidad del suavizante al seleccionar un número pequeño de funciones base, comúnmente menor a 10; el enfoque es iniciar con las n funciones base y penalizarlas de tal manera que sea flexible el ajuste. El problema se reduce en lo siguiente:

Sea $\Sigma = W^{-1} = diag(w_1^{-1}, \ldots, w_n^{-1})^T$ una matriz de pesos *a priori*, entonces el criterio de mínimos cuadrados penalizados está definido por

$$S(f) = (y–f)^T \Sigma^{-1}(y - f) + \lambda f^T K f$$

donde:

- K representa la matriz de penalización por rugosidad, en ella se encuentra la combinación en pares de las funciones base, definidas usualmente por el término de una derivada o por algún operador diferencial aplicado a esta función base.

- λ representa el parámetro del suavizamiento, es importante resaltar que cuando $\lambda \to \infty$ entonces $f^t K f \to 0$ y esto podría implicar que se está trabajando con una aproximación lineal.

El tercer método, **regresión local**, se enfoca en ajustar curvas mediante suavizamientos utilizando únicamente observaciones cercanas a x_0, el método se puede resumir en:

- Para cada x_0, el entorno, para elegir el tamaño de la ventana se utiliza el parámetro de suavizamiento, que en este caso representa la proporción de observaciones que se utilizaran en la regresión local.

- Se asigna pesos a los vecinos más cercanos, considerando que los valores más cercanos a x_0 tendrán un peso mayor que los más lejanos.

- Finalmente la regresión ponderada, se hace sobre el conjunto de las $x's$ pertenecientes a la vecindad alrededor de x_0.

El cuarto método de suavizamiento, **el vecino más cercano**, se describe de la siguiente manera: primero consideremos la figura 3.1 y supongamos que la elipse morada representa un scatterplot entre x y y, además tomemos x_0 donde $\hat{y_0}$ representa la esperanza condicional

calculada. Una forma de calcular esta $\hat{y_0}$ es mediante los valores de x cercanos a x_0; en la figura el rectángulo representa el área, vecinos más cercanos, generando así una vecindad alrededor de x_0. En primera instancia la esperanza condicional es un buen resumen de y, tendrá menor sesgo y será más estable contra la esperanza considerando únicamente a x_0. Para más detalle revisar [Berk, 2008].

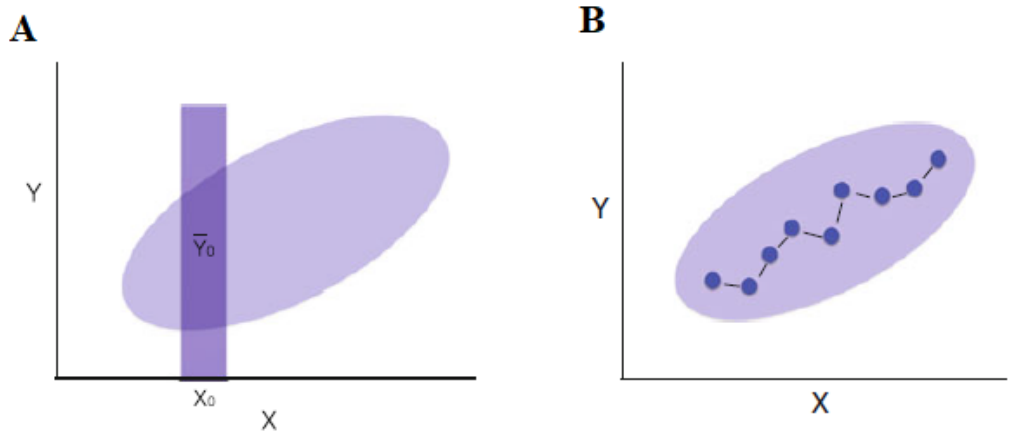


Figura 3.1: **A.** Vecino más cercano. **B.** Interpolación de Esperanzas condicionales.

El principal problema de este método es cómo definir el tamaño de las vecindades y la cantidad de vecinos en ellas, existen varias opciones para ello, por ejemplo, tomar los k-más cercanos a x_0 o el $k\,\%$ más cercano a x_0, e incluso es posible generar una combinación de ambas formas tomando en cuenta siempre la densidad de y y no es necesario cumplir la simetría en las vecindades. Entre mayor sea el tamaño de la vecindad el valor de la variable respuesta cambiará y el suavizamiento será menos variable; por el contrario, vecindades pequeñas generarán estimaciones menos sesgadas, menos estables y más variables.

3.2. Splines

Un spline es una función polinómica por partes obtenidas mediante la partición del rango de X en k regiones y la unión de cada una de estas regiones es llamada nodo. De igual forma, los splines son usados regularmente como herramienta en el suavizamiento de curvas, como se vio en la sección anterior y además proporcionan una transición entre un modelo paramétrico de regresión y un modelo totalmente suavizado. Estos splines o particiones dependen de tres elementos:

- Grado del polinomio.
- Número de nodos.
- Localización de los nodos.

De acuerdo con [Yee, 2015], el tipo de spline más popular es aquel con polinomio de grado tres y con la existencia de las primeras dos derivadas, además de ser continuas en los

nodos, para así, al unirlos, finalizar con una función continua y suavizada, como se muestra a continuación.

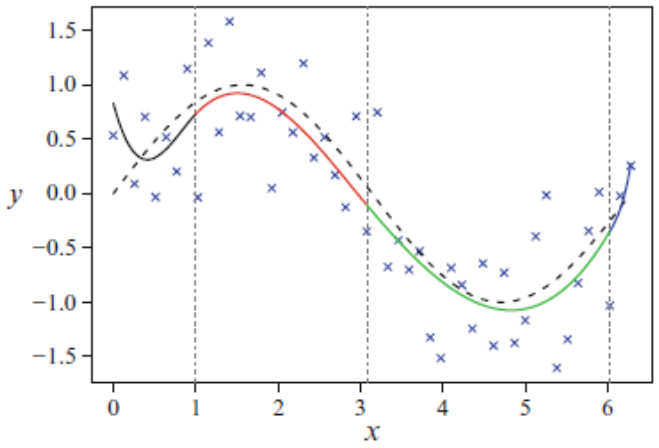


Figura 3.2: Ejemplo de una función calculada mediante splines.

3.2.1. Splines cúbicos

Los splines cúbicos dependen de los nodos, por lo que es importante la cantidad y la localización, mismos que son ubicados usualmente en los cuantiles de x y el tamaño es seleccionado entre [3, 7], siempre considerando el tamaño de la base de datos con el que se está trabajando y el comportamiento de los datos.

Este tipo de spline, al ser utilizado en un suavizamiento, se convierte en un problema de optimización, por ello se debe buscar la solución que minimice la penalización de la suma de los siguientes mínimos cuadrados:

$$\underbrace{\sum_{n=1}^{\infty} (y_i - f(x_i))^2}_{1} + \underbrace{\lambda \int_a^b (f''(x_i))^2 \, dt}_{2}$$

La primera parte explica la relación entre la variable observada y su estimado, mientras que la segunda parte de la ecuación penaliza la curvatura de la función, además contiene a λ el parámetro de suavizamiento.

3.2.2. B-Splines

Los splines base o B-splines son funciones compuestas por la unión de varios polinomios conectados entre sí.

Un B-spline de grado q se puede reducir a lo siguiente:

- Contar con $q+1$ polinomios de grado q. Una característica de estos polinomios es que cuentan con la $(q-1)$-ésima derivada continua en los puntos de unión.
- Se unen en p nodos internos.
- Para cada valor de x, existen $q+1$ B-splines no nulos.

Por ejemplo, en la imagen presentada a continuación se muestran algunos ejemplos de B-splines cuando:

- $q = 1$ Es similar a una función de rectángulos.
- $q = 2$ Spline lineal, se puede mapear como una función lineal en los splines, dado que q=2, la función está obligada a ser continua.
- $q = 3$ Spline cuadrático, con la primera derivada continua en los nodos.
- $q = 4$ Spline cúbico, mencionado anteriormente.

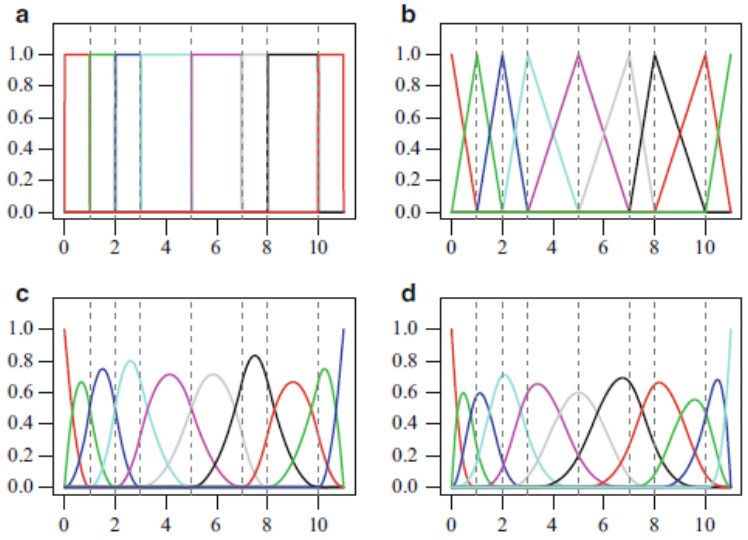


Figura 3.3: Ejemplo de una funcion calculada mediante splines [Yee, 2015].

Una de las características de los $B - splines$ es que tienen un soporte mínimo, lo cual implica que, al generar las vecindades, los coeficientes de un $B_i - spline$ están relacionados en menor medida con los coeficientes de $B_{i+1} - spline$, lo que se traduce como una cantidad de valores sobrepuestos mínima.

3.2.3. P - Splines

Los p-splines o splines con penalización utilizan menos parámetros que los splines de suavizado, a su vez la selección de los nodos no es determinante como en el caso de los splines de regresión.

La metodología de los p-splines se puede resumir a continuación:

- Utilizar una base para la regresión.
- Modificar la función de verosimilitud, lo cual se lleva a cabo mediante una penalización basada en diferencias entre coeficientes adyacentes.

La base para la regresión se puede calcular de varias maneras, las dos principales son las bases polinomios truncados y los utilizados en los B-splines.

3.3. Modelos Aditivos Generalizados

Los modelos aditivos generalizados (GAM) son una extensión no paramétrica de los modelos lineales generalizados y una generalización de un modelo de regresión común. Usualmente un modelo aditivo con p predictores es expresado de la siguiente manera:

$$g(\mu(x_i)) = \eta_i = \beta_i + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i$$

Siendo así una suma de funciones, donde cada f_j es el suavizamiento de la covariable correspondiente, x_j. Estos modelos pierden la linealidad y dan mayor flexibilidad al utilizar los suavizamientos, de esta manera se permite un mejor ajuste, sin embargo, al asumir la aditividad, se mantiene la relación entre las covariables con el fin de evitar la pérdida de información.

De la misma manera que un modelo lineal generalizado, este tipo de modelos permite utilizar diferentes funciones liga o aproximar más de una forma de distribución. En el caso de una regresión, los coeficientes ayudan a escalar la función de los predictores, aquí ese rol no puede ser distinguido del rol de la transformación dado que no son identificables.

Para ajustar un modelo aditivo existen diferentes formas en la formulación y en la estimación, la diferencia radica en el uso y las combinaciones de los tipos de suavizamientos aplicados individualmente en cada covariable. En seguida se muestran algunos ejemplos de técnicas a utilizar en modelos aditivos, considerando también aquellos modelos semi-paramétricos; para más detalle revisar [Hastie y Tibshirani, 1990]:

- Regresión lineal múltiple. En este caso se puede dar un enfoque al transformar las variables paramétricamente con un logaritmo y raíz cuadrada, entre otras. De igual forma, se puede considerar cada variable y generar un conjunto de polinomios, éstos representados por f_j en la ecuación anterior.

- Funciones bases. Aquí se utiliza un conjunto de funciones base y se generan múltiples modelos de regresión; en este método los splines de regresión definen un conjunto especial de funciones base, los cuales son aplicados parcialmente generando un único polinomio, igualmente, son altamente ligados a los B-splines. Se busca una función tal que para $i = 1, \cdots, n$
$$y_i = f(x_i) + \epsilon_i$$

- Splines de suavizamiento.

- Modelo general. Éste permite estimar cada función con el suavizamiento conveniente o con el que mejor se ajuste, entre las selecciones están los suavizamientos con splines cúbicos, los suavizamientos con pesos predeterminados y aquellos de núcleo.

El algoritmo de la estimación de la función y de α se puede resumir como se muestra a continuación, para ello se considera $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\hat{f}_j \equiv 0$, para todo i, j.

1: **for** $i = 1, \cdots, n; j = 1, \cdots, p$ **do**
2: $\quad \hat{f} \approx S_j[y – \hat{\alpha} - \sum_{k \neq j} \hat{f}_{ij}]$
3: $\quad \hat{f}_j - \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{ij} \rightarrow \hat{f}_j$
4: **end for**

Dado que existen diferentes maneras de calcular el ajuste de un modelo aditivo generalizado, en el siguiente listado se describen brevemente las características de la estructura que se puede esperar que tenga un efecto en el desempeño de los procedimientos de ajuste.

- **Cantidad de información.** Esta característica puede tener un efecto en el desempeño del modelo, para ello podemos medir el radio señal-ruido.
- **Tipo de respuesta.** Si es continua o discreta, y cómo afecta la estimación realizada, en caso de no ser tan evidente la distribución, podría afectar considerablemente al desempeño.
- **Número de covariables no influyentes.** Al contar con variables no significativas para el modelo, y de no ser posible quitarlas, esto podría traer problemas al desempeño del modelo, problemas derivados de no sólo estimar un parámetro, sino una función de la covariable.
- **No linealidad en los datos.** Se busca medir un procedimiento que ofrezca más complejidad, por lo que debería presentar medios para la regulación automática de la complejidad, de esta manera, se podría ajustar a un modelo complejo sólo cuando sea necesario y a uno simple cuando sea suficiente.

Hasta este punto se ha presentado una gama variada de modelos para variables binarias, por lo que, una vez concluida la sección teórica, lo siguiente es aplicar los conocimientos adquiridos. El objetivo central es aproximar un modelo binario con un modelo aditivo generalizado, sin dejar de lado los modelos lineales.
En el proceso de modelación existen diferentes metodologías como KRISP, SEMMA, KDD entre otras, a continuación se utiliza una metodología KRISP, la cual consiste en entendimiento de negocio o entendimiento del problema, entendimiento de los datos, transformación de los datos, modelación y evaluación, derivado de lo anterior, se cuentan con las herramientas necesarias para comenzar este proceso.

Capítulo 4

Parkinson

La enfermedad de Parkinson es una condición neurodegenerativa asociada con síntomas neuropsicológicos, sensoriales, vocales y afectaciones en el sistema motor [Dias *et al.*, 2016]. Asimismo, muestra transtornos de voz y del habla que se encuentran presentes en aproximadamente el 90 % de las personas que padecen esta enfermedad, por ello la detección temprana de esta afectación ayuda a la selección de un tratamiento.

Debido a que una de las principales características de este padecimiento son las alteraciones en el habla, las grabaciones de voz han sido consideradas como marcadores biológicos en el diagnóstico de este tipo de enfermedades; puesto que el Parkinson afecta desde su aparición, el daño puede no ser perceptible hasta etapas más avanzadas; en ese sentido, el análisis de voz puede ser considerado como un estudio preventivo y reactivo.

El seguimiento que se le dará a la información recolectada está basada en la detección de Parkinson, para ello se cuenta con la información de un grupo de 80 individuos donde el 50 % cuenta con la enfermedad y el otro 50 % no la tiene.

El análisis se llevará a cabo en grabaciones de voz del grupo de personas, usualmente los parámetros de frecuencia de voz (Jitter) y la amplitud de ésta (Shimmer) son usados en la estudio de grabaciones de voz (ver figura 4.1). Ambos parámetros de perturbación son esenciales en el análisis de grabación de voz, mismos que se obtienen por prolongaciones largas de las cuerdas vocales. Las perturbaciones en la voz se ven reflejadas en el Jitter debido a que las grabaciones de voz en pacientes con alguna patología usualmente presentan valores más grandes en éste, ya que es afectado principalmente por la pérdida del control de la vibración de las cuerdas vocales. Para el Shimmer, éste cambia con la reducción de la resistencia glotal y una lesión grave en las cuerdas vocales porque están correlacionados con la presencia de ruidos y de la respiración [Teixeira *et al.*, 2013].

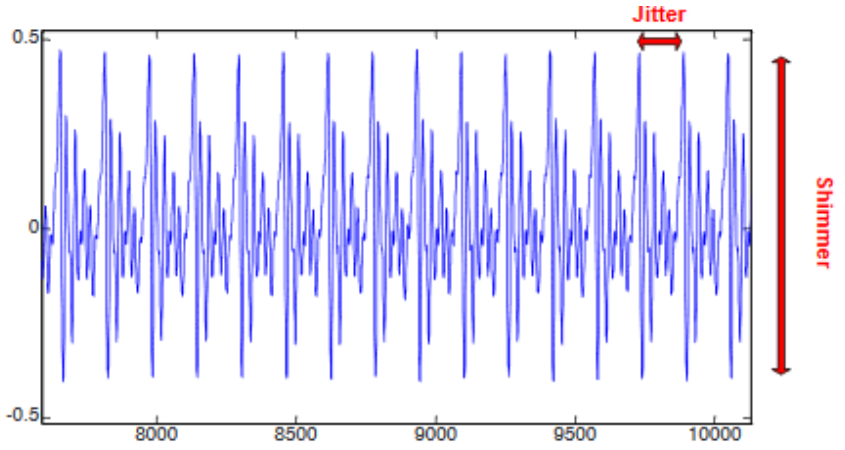


Figura 4.1: Ejemplo de las perturbaciones en Shimmer y Jitter.

4.1. Entendimiento de los datos

En esta sección se explica el funcionamiento de las grabaciones y su transformación a vectores, donde cada uno representa la información de cada individuo en un tiempo t_i. Se obtuvieron 3 grabaciones de 80 individuos; cada grabación de voz fue procesada y transformada para obtener 44 características, obteniendo así un total de 240 observaciones clasificadas de la siguiente manera:

- Medidas del tono:
  - Jitter relativo.
  - Jitter absoluto.
  - Perturbación relativa promedio (RAP, por sus siglas en inglés).
  - Cociente de las perturbaciones en el tono (PPQ, por sus siglas en inglés).
- Medidas de amplitud:
  - Shimmer local.
  - Shimmer en dB.
  - Amplitud de 3 periodos (APQ3, por sus siglas en inglés).
  - Amplitud de 5 periodos (APQ5, por sus siglas en inglés).
  - Amplitud de 11 periodos (APQ11, por sus siglas en inglés).
- Medidas de la relación armónica:
  - HNR05: 0 - 500 Hz.
  - HNR15: 0 - 1500 Hz.
  - HNR25: 0 - 2500 Hz.
  - HNR35: 0 - 3500 Hz.

    - HNR38: 0 - 3800 Hz.

- Coeficientes Cepstrales en las Frecuencias de Mel:
    - MFCC0 - MFCC12.
    - Sus derivadas respectivamente, Delta0 - Delta12.
- Entropía de densidad del periodo de recurrencia (RPDE, por sus siglas en inglés).
- Análisis de flutación sin tendecia (DFA, por sus siglas en inglés).
- Entropía del periodo de tono (PPE, por sus siglas en inglés).
- Relación de excitación glotal al ruido (GNE, por sus siglas en inglés).

La base de datos [Pérez, 2016] además cuenta con el identificador para cada individuo, una variable de género y el estatus del mismo respecto a la enfermedad.

Al hacer un análisis sobre las variables, se observa que cada una se distribuye de manera particular, a continuación se muestra a detalle cada una de las variables estadísticamente hablando.

Las variables que nos ayudan a medir el tono de voz, Jitter, se encuentran en valores muy cercanos a cero (ver figura 4.2). El Jitter relativo es la variable que toma los valores más altos, mientras que el Jitter absoluto es aquella que toma valores más pequeños; en caso de requerir una selección de variables se podría considerar únicamente el Jitter relativo.

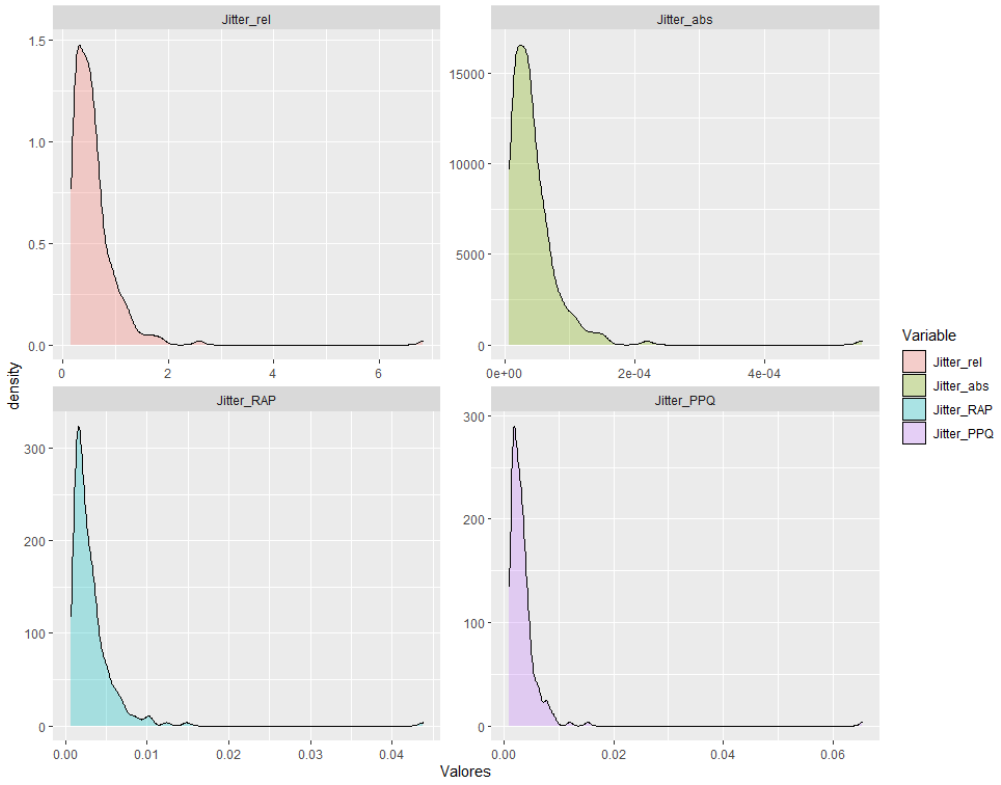


Figura 4.2: Presenta un resumen de la distribución de variables relacionadas al Jitter.

Al analizar los boxplots correspondientes al Jitter (ver figuras 4.3 y 4.4), se observa que las cuatro variables cuentan con un outlier entre sus valores, el cual podría causar sesgo en la parte de modelación. Del mismo modo, la distribución de las cuatro variables es bastante similar, la diferencia se presenta en cuanto al intervalo donde cada una se ve reflejada.

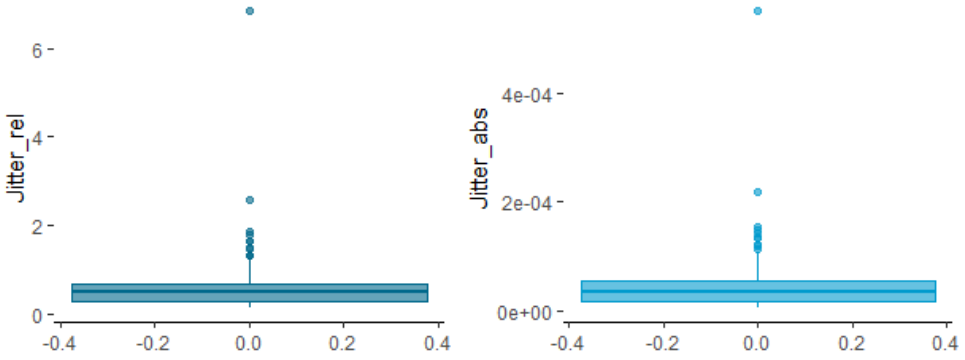


Figura 4.3: Boxplot de variables relacionadas al Jitter.

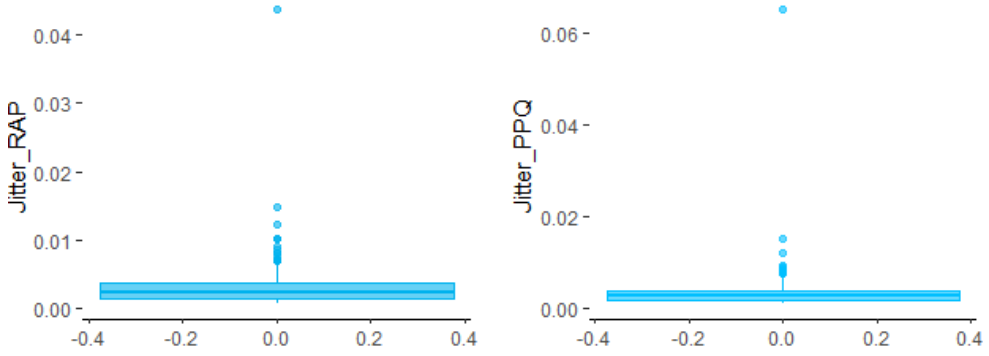


Figura 4.4: Boxplot de variables relacionadas al Jitter.

Las variables que nos ayudan a medir la amplitud de las ondas, Shimmer, toman valores en el intervalo (0, 0.1); en su mayoría toman valores cercanos al 0.01. En cuanto las variables asociadas al Shimmer, las variables AP3, APQ5 y APQ11, se puede notar el aumento conforme crece el número de ondas que se están comparando, como se observa en la gráfica de la figura 4.5. La distribución de las variables Shim_APQ3 y Shim_APQ11 es similar y son las que toman los valores más pequeños al encontrarse más hacia la izquierda.

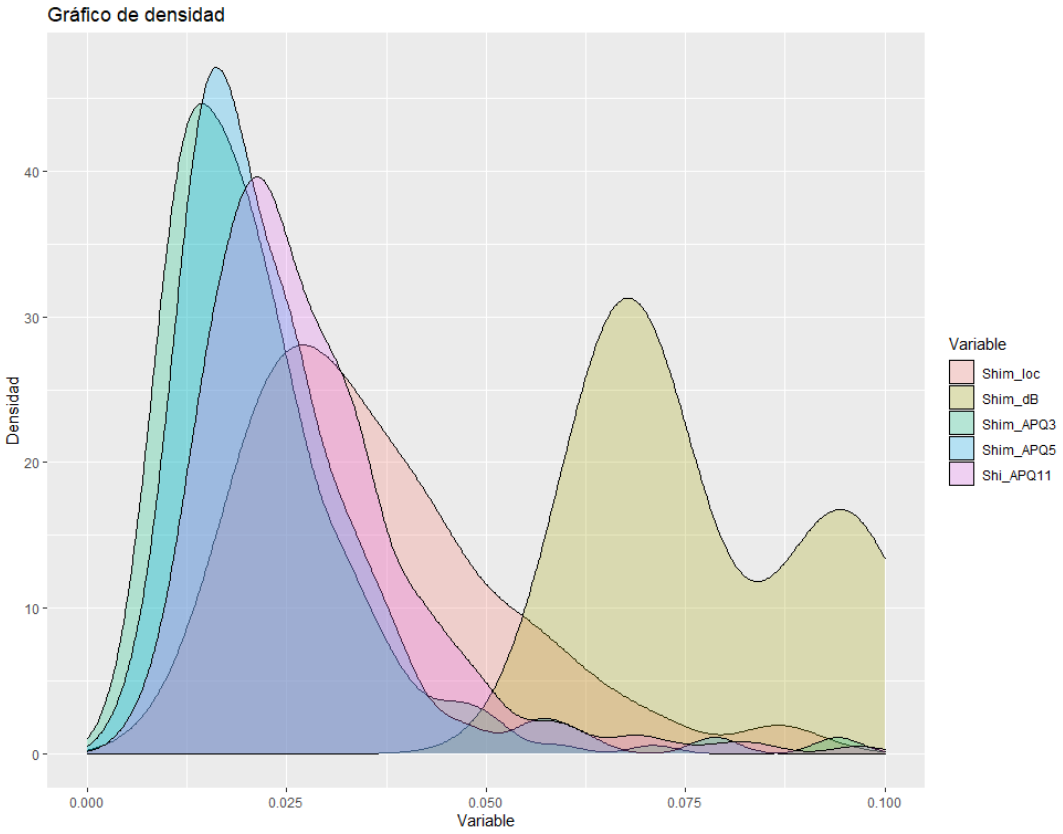


Figura 4.5: Gráfica de densidad de variables asociadas al Shimmer.

Respecto a las variables de medición armónica, no se presenta una vasta diferencia entre los hertz (Hz) consecutivos, pero sí se observa una divergencia considerable en el comportamiento entre HNR05 y HNR38, a saber: por un lado, las mediciones menores a 1500 Hz toman en su mayoría valores más pequeños y su distribución tiende a la izquierda, por otro lado, las variables de 3500 Hz y de 3800 Hz ocupan valores más grandes y sus distribuciones se aproximan más hacia la derecha, mientras que el caso de 2500 Hz tiene una distribución más centrada, la cual se parece a una distribución normal; véase en la figura 4.6.

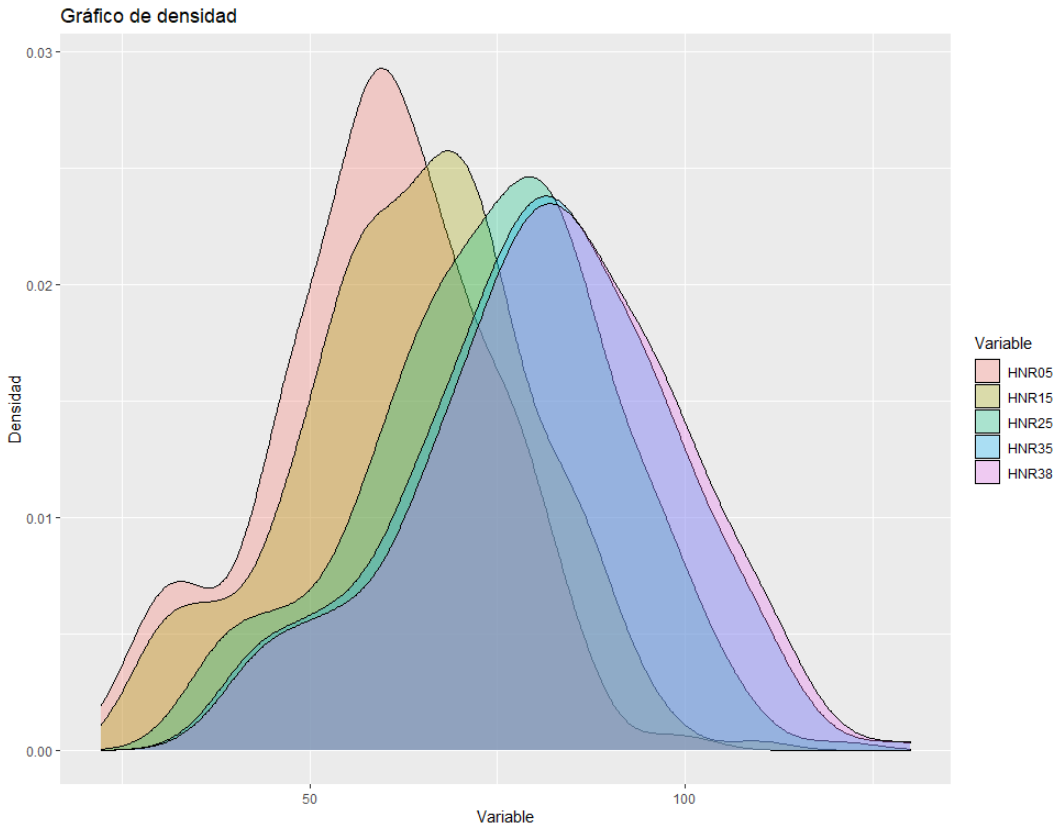


Figura 4.6: Gráfica de densidad de variables HNR.

Ahora bien, las variables RPDE, DFA, PPE y GNE siguen distribuciones particulares, esto es: la relación entre la excitación glotal y el ruido es la variable con menor varianza y la que acapara los valores más cercanos a 1, por su parte, las distribuciones de las variables RPDE y DFA se encuentran centradas y en el caso de PPE la distribución está ligeramente sesgada a la izquierda; todo lo antes detallado se advierte en el diagrama de caja de la figura 4.7.

Por otra parte, los Coeficientes Cepstrales de las Frecuencias de Mel (MFCC) muestran, según el instante de análisis, las características locales de la señal de voz asociadas al tracto vocal, en suma, estas trece variables tienen una distribución bastante similar, al igual que sus deltas hacen lo propio, como se puede observar en la figura 4.8.

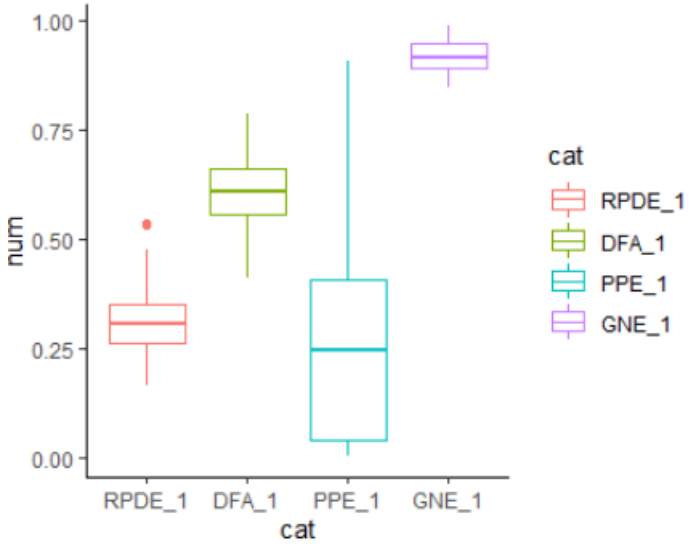


Figura 4.7: Boxplot de distribución de las variables RPDE, DFA, PPE y GNE.

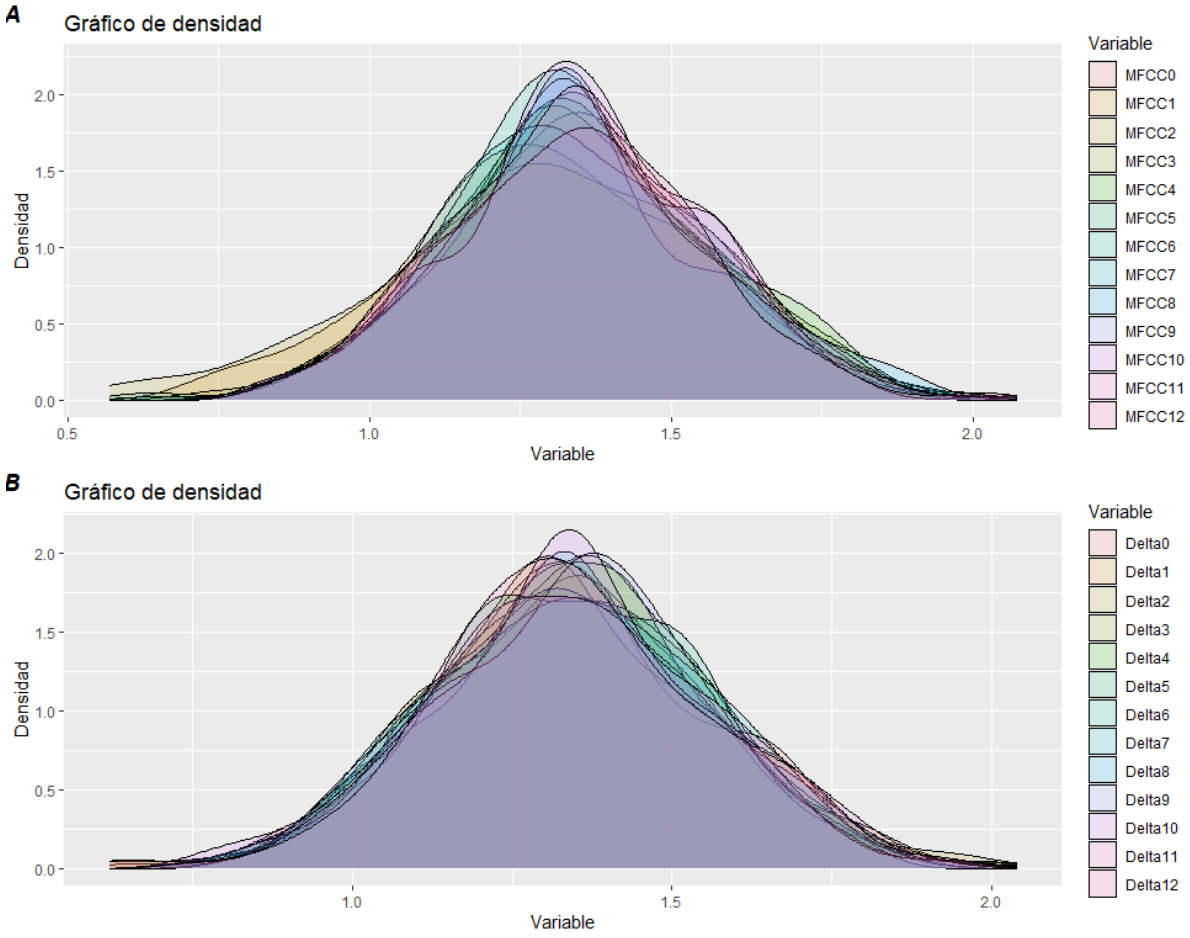


Figura 4.8: **A.** Densidad de los Coeficientes Cepstrales de las Frecuencias de Mel. **B.** Gráfica de densidad de las derivadas de los Coeficientes Cepstrales de las Frecuencias de Mel.

Tras un breve análisis estadístico sobre la distribución de las variables numéricas y la descripción de éstas —el cual se realizó mediante el cálculo de la correlación de Pearson entre dos variables—, es pertinente concluir que por la naturaleza de las mismas, existe gran correlación entre ellas; en la gráfica de la figura 4.9 se representan con color verde aquellas variables correlacionadas positivamente y con color azul aquellas cuya correlación es negativa.

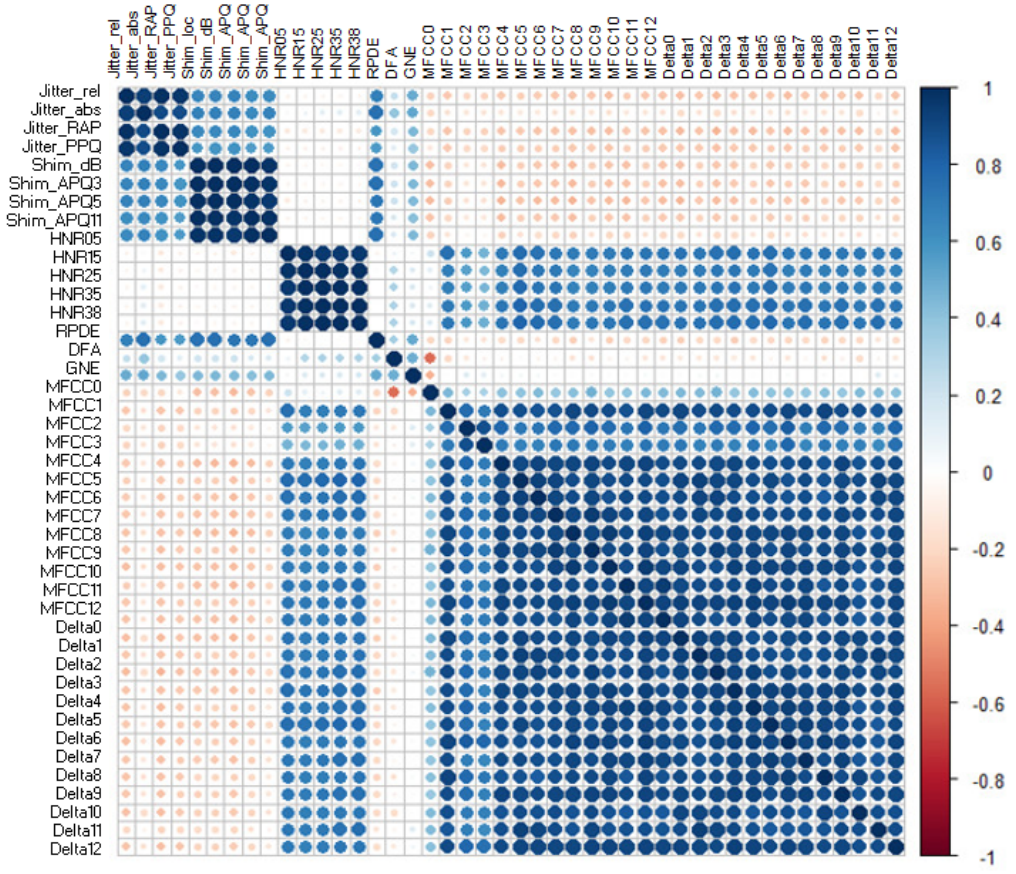


Figura 4.9: Gráfico de correlación de Pearson de las variables numéricas.

Como se puede reparar, las variables se agrupan de acuerdo a la correlación que existe entre ellas y, además, cabe resaltar que ésta es positiva; es decir, si una de las variables del grupo i crece, el resto también lo hará. Los grupos identificados son los siguientes:

- Grupo 1:
  - Jitter_rel, Jitter_abs, Jitter_RAP y Jitter_PPQ.
- Grupo 2:
  - Shim_loc, Shim_dB, Shim_APQ3, Shim_APQ5, Shi_APQ11 y RPDE.
- Grupo 3:
  - HNR05, HNR15, HNR25, HNR35 y HNR38

- Grupo 4:
    - MFCC0 a MFCC12 y Delta0 a Delta12.

Asimismo, con base en el análisis de correlación se ratifica que también el Grupo 1 y el Grupo 2 se corresponden, pues el Jitter representa la frecuencia y el Shimmer la amplitud en las grabaciones.

4.2. Preparación de los datos

4.2.1. Tratamiento de valores extremos

En el apartado pasado se describió el comportamiento de las variables y se identificó el comportamiento extraño de un individuo a quien se le trata de manera especial. De acuerdo a lo expuesto en el capítulo 1, a continuación se utilizará el paquete de *Amelia* en R (creado para el tratamiento de valores ausentes), el cual aplica el algoritmo mencionado anteriormente.

A propósito, es oportuno enfatizar que para identificar el ID del valor extremo, "PARK-18-2", debe hacerse conforme a las variables Jitter_abs, Jitter_rel, Jitter_RAP y Jitter_PPQ. Dicho lo anterior, el proceso para la sustitución del valor extremo es el siguiente:

1. Identificar el valor extremo, "PARK-18-2".
2. Sustituir los valores del Id "PARK-18-2" en las variables Jitter_abs, Jitter_rel, Jitter_RAP y Jitter_PPQ por valores ausentes.
3. Aplicar la función *Amelia*, especificado que la base de datos cuenta con variables categóricas (Gender) y un identificador (ID).
4. Seleccionar una base de datos resultante de la simulación, dadas las circunstancias es indistinto tomar alguno de los cinco *datasets* generados.

Una vez graficada la densidad de la simulación, como se observa en las figuras 4.10 y 4.11 (en ambos casos en azul figura la densidad original y con rojo se grafica la densidad simulada), en virtud de que, en este caso, es un único valor el imputado, sobresale en color rojo el nuevo valor que tendrá:

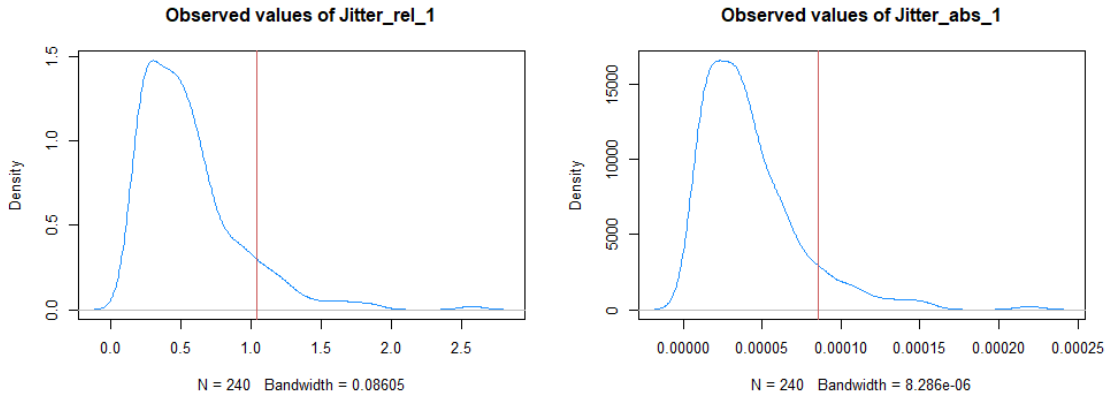


Figura 4.10: Densidad para las variables Jitter_rel y Jitter_abs de acuerdo a la simulación de la función Amelia.

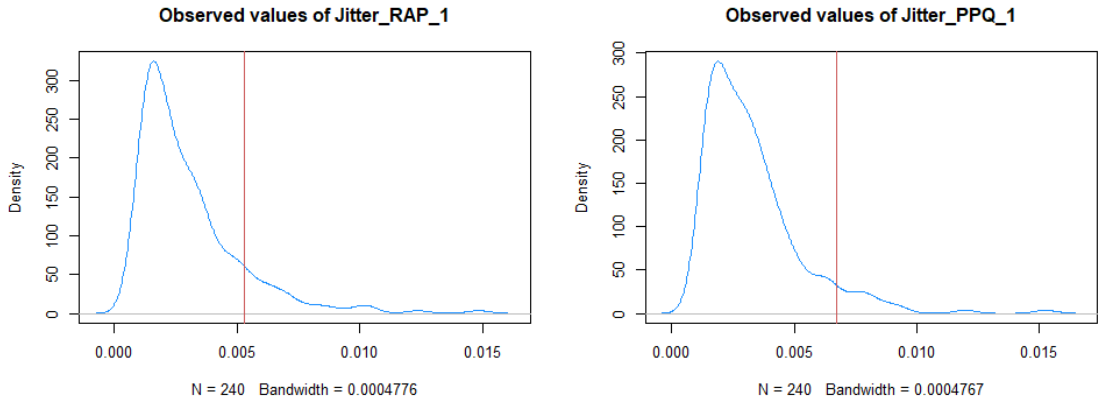


Figura 4.11: Densidad para las variables Jitter_RAP y Jitter_PPQ de acuerdo a la simulación de la función Amelia.

Finalmente, en los diagramas de la figura 4.12 se comparó la densidad de las variables simuladas con las variables observadas; en éstos se descubre que en las cuatro variables la distribución simulada se apega a la distribución original, a partir de lo cual se concluye que se encuentran cercanas a la identidad. Referente al caso de Jitter_abs, al ajustar la cola de la distribución no se logra un ajuste tan preciso como en el resto de los valores donde la simulación encaja bastante bien, por lo que al valor imputado final de Jitter_abs se le puede considerar como una buena aproximación.

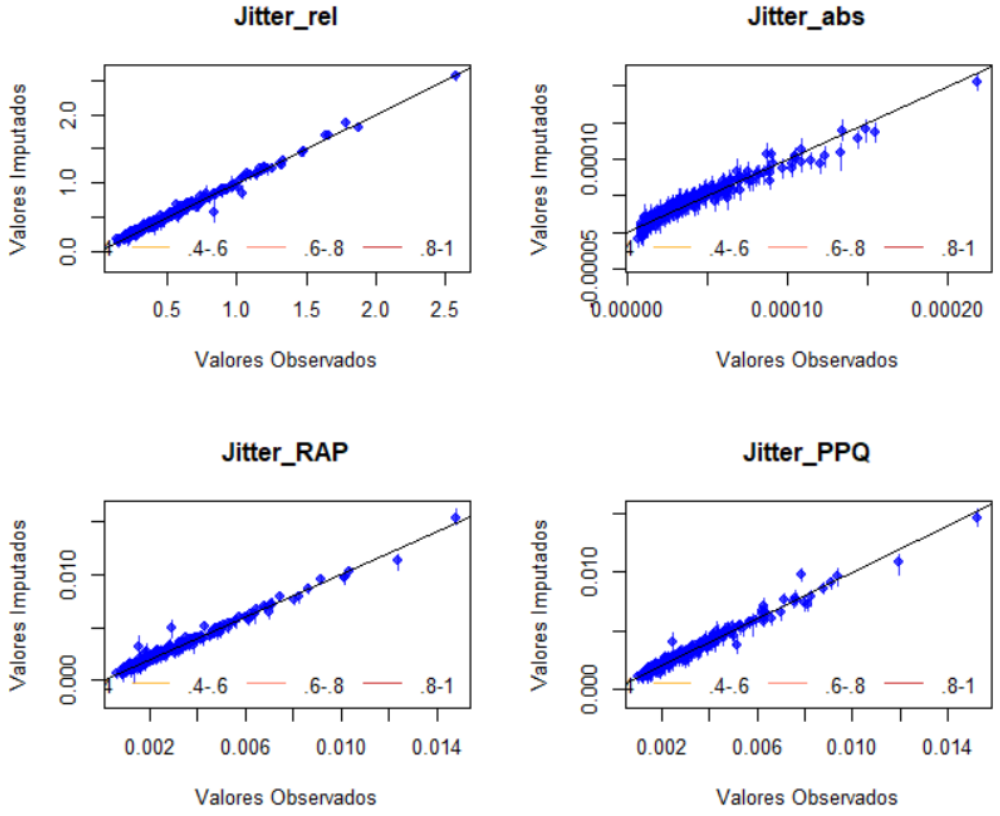


Figura 4.12: Scatterplot del valor observado contra el valor imputado.

4.2.2. Reducción de dimensiones

Toda vez que la correlación existente entre las variables y que uno de los supuestos de los modelos de regresión generalizados y los modelos aditivos generalizados es la independencia de las variables y de los valores observados, en adelante se efectúa el análisis de componentes principales de suerte que se genera una variable, la cual es combinación de las variables correlacionadas (pertenecientes a los grupos generados anteriormente) y ésta la sustituirá.

Así pues, tras el análisis de los componentes principales en el grupo 1 —el cual está constituido por las variables imputadas Jitter_rel, Jitter_abs, Jitter_RAP y Jitter_PPQ—, se concluye que el primer componente (Jitter_rel) es, por una parte, el que explica el 100 % de la varianza de los datos originales y, por otra parte, el que será la nueva variable sustitutiva de las del grupo 1, quedando compuesta, así, por las variables Jitter_rel y Jitter_abs, ver figura 4.13.

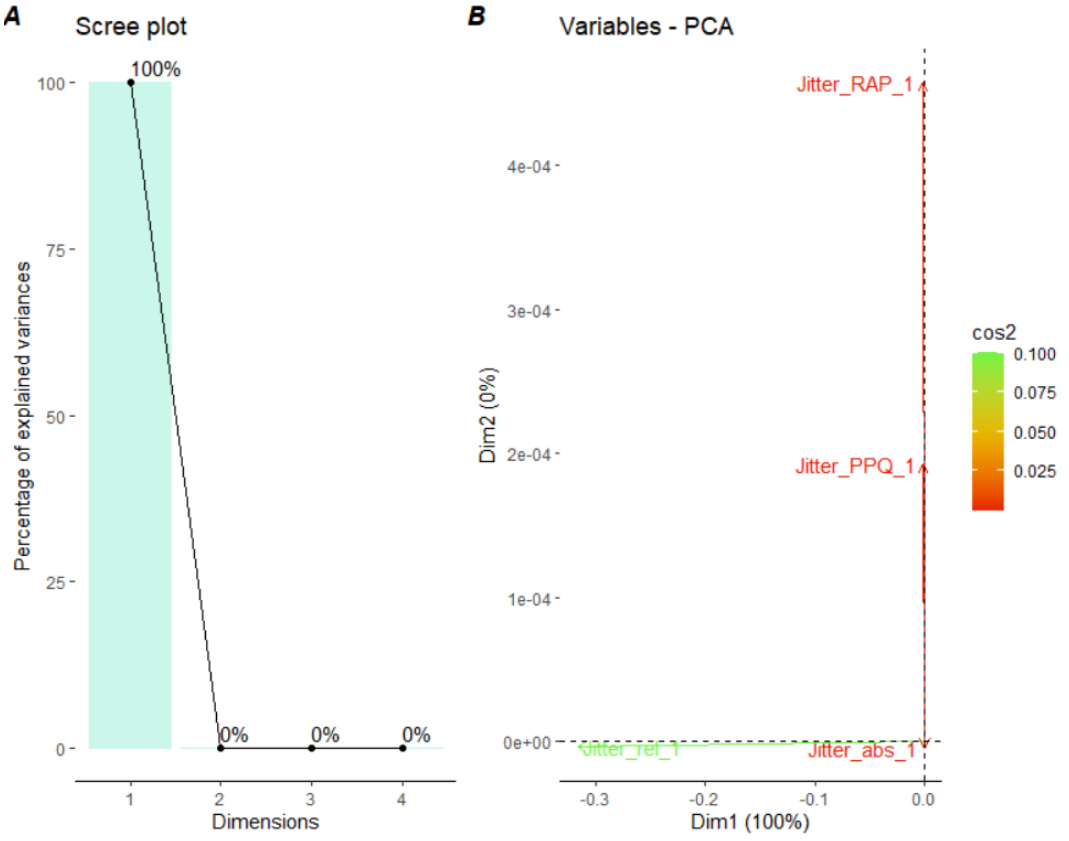


Figura 4.13: **A**. Varianza explicada por componente principal. **B**. Correlación entre variables originales y las primeras dos componentes.

Por otro lado, en el grupo 2 se concentran las variables Shim_loc, Shim_dB, Shim_APQ3, Shim_APQ5 y Shim_APQ11. Al respecto, cabe señalar que, debido a la alta correlación existente entre las variables, la combinación de éstas explica el 99.9 % de la varianza total de los datos originales, mientras que el resto de los componentes no agrega una cantidad significativa de varianza; por lo anterior, se tomará sólo el primer componente como la variable sustitutiva.

Ahora bien, la variable más correlacionada con el primer componente es Shim_dB, en cambio, el resto de las variables lo están con el segundo componente, como se observa en **B** de la figura 4.14.

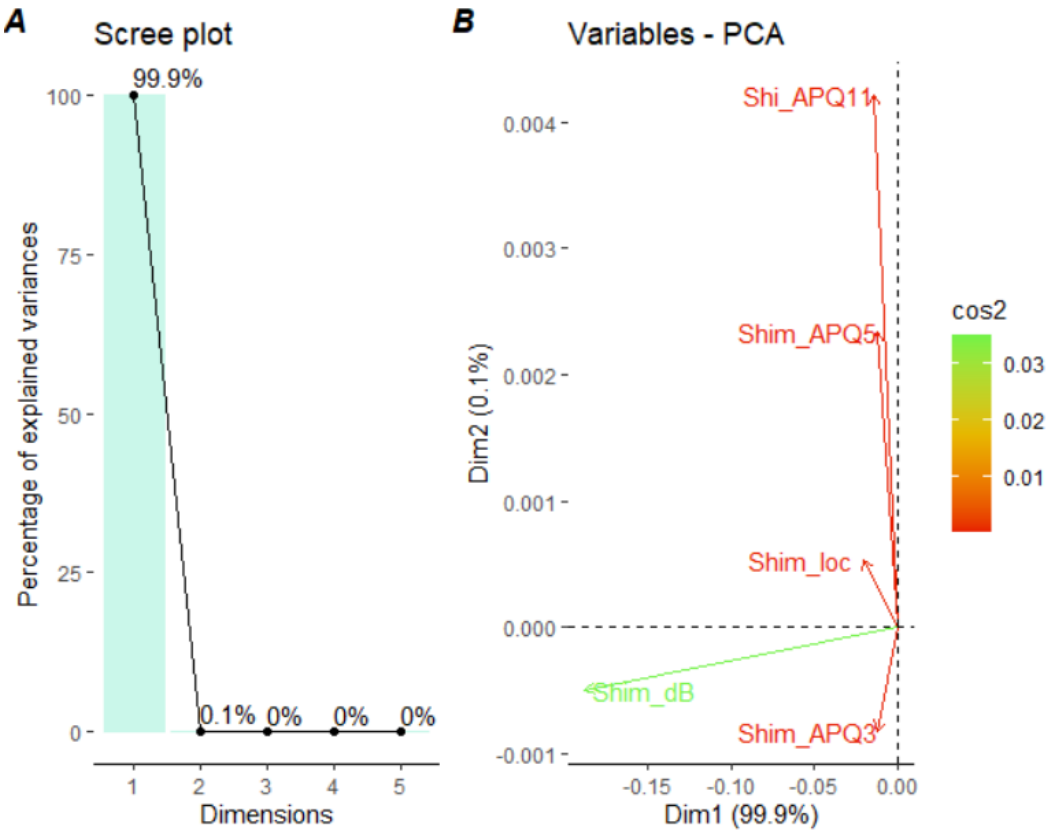


Figura 4.14: **A**. Varianza explicada por componente principal. **B**. Correlación entre variables originales y las primeras dos componentes.

En lo que concierne al escrutinio del tercer grupo —cuyos componentes son HNR05, HNR15, HNR25, HNR35 y HNR38—, se hace patente que el primer componente explica el 98.6 % de la varianza total de los datos originales, lo cual es un valor significativo que coadyuva a reconocer dicho componente como el sistitutivo de las variables de este grupo. Además, gracias al gráfico **B** de la figura 4.15, el cual se exhibe a continuación, se pondera que las variables HNR15 y HNR25 se correlacionan mayormente con el primer componente, mientras que HNR05 con el segundo componente.

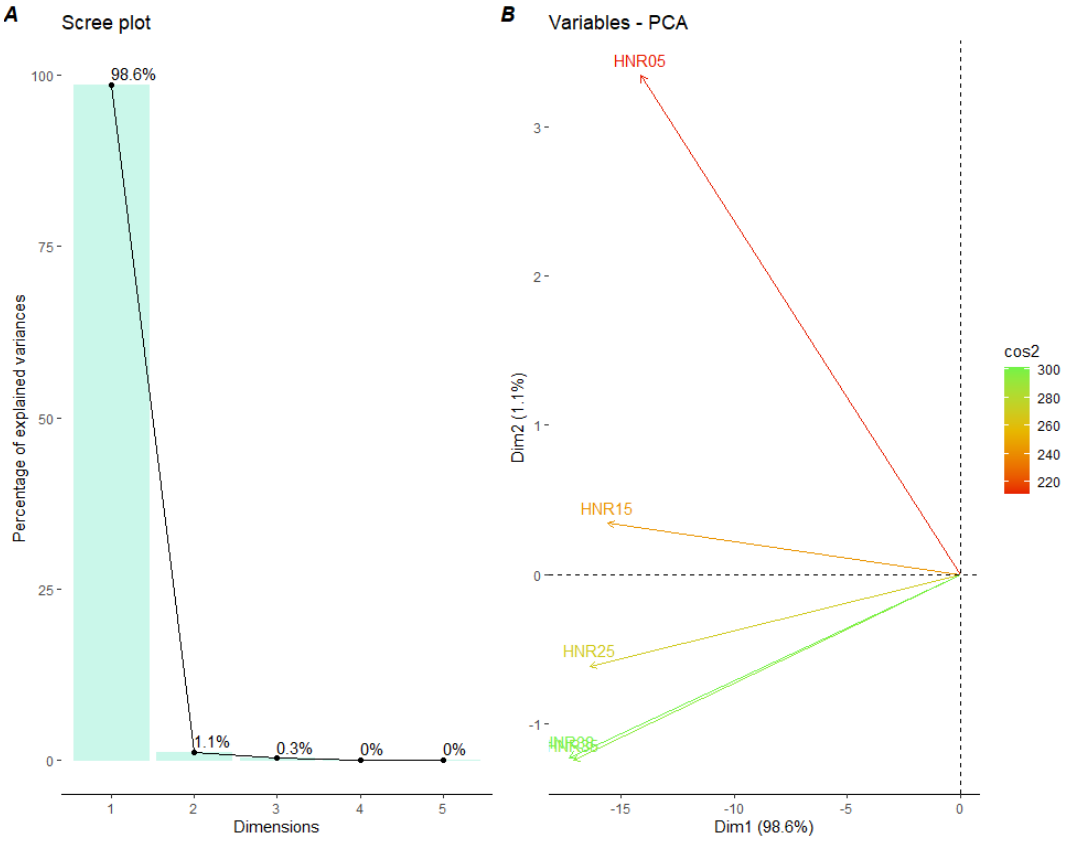


Figura 4.15: **A**. Varianza explicada por componente principal. **B**. Correlación entre variables originales y las primeras dos componentes.

Por último, el cuarto grupo integrado por 26 variables, MFCC0 a MFCC12 y sus respectivas derivadas Delta0 a Delta12, es el grupo más grande de los que se han analizado. En virtud de la naturaleza de las variables, la mayoría de éstas se correlacionan con el primer componente; en cambio, las pocas variables correlacionadas principalmente con el segundo componente son MFCC1 y MFCC2. En suma, en este grupo el primer componente explica sólo el 73.8 % de la varianza original, como se revela en la figura 4.16; si bien podrían tomarse en cuenta los primeros dos componentes —ya que 24 de las 26 variables se encuentran cercanas al primer componente, véase la imagen B de la figura 4.16—, las variables se sustituirán únicamente por aquella (MFCC1).

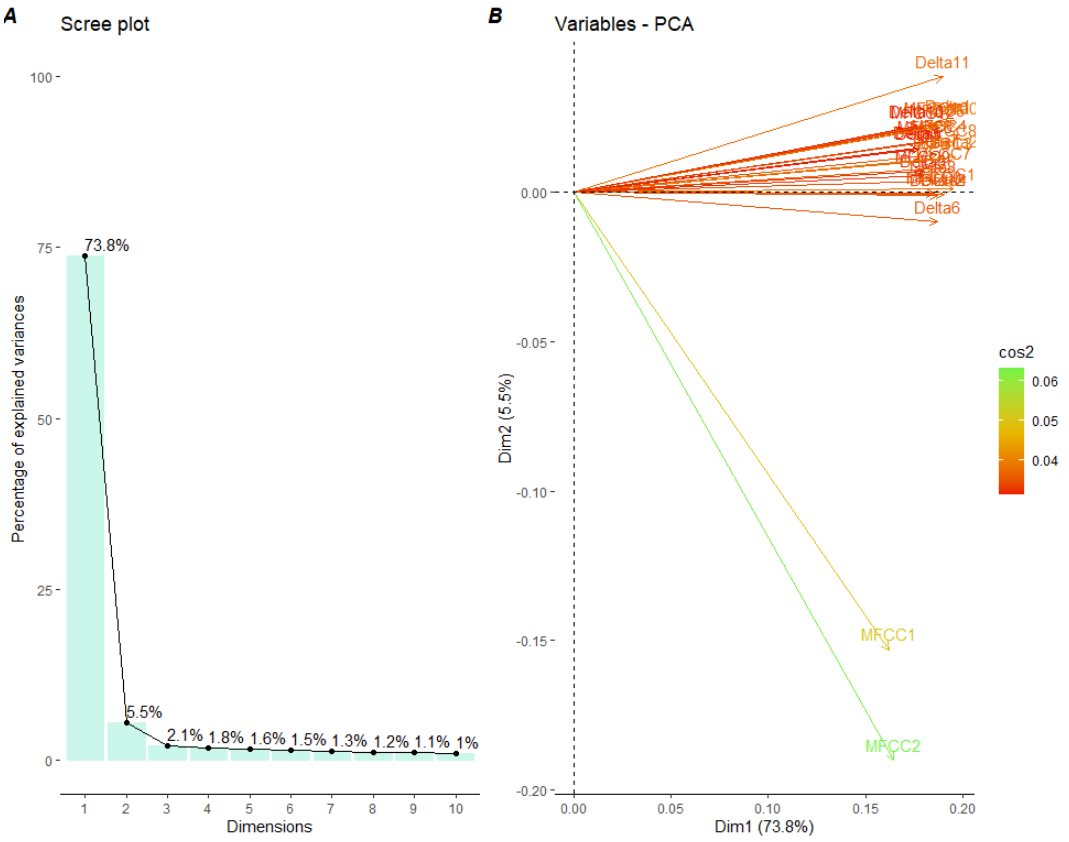


Figura 4.16: **A**. Varianza explicada por componente principal. **B**. Correlación entre variables originales y las primeras dos componentes.

Finalmente, después de las transformaciones ponderadas mediante los análisis y, a su vez, efectuadas, se generó una base de datos con las siguientes variables:

- ID: identificador del sujeto.
- Status: variable original y variable objetivo, a saber: el valor de 0 atañe al paciente sin la enfermedad y 1 cuando el paciente padece la enfermedad.
- Gender: variable original, sin ir más lejos: el valor de 0 corresponde al paciente masculino y 1 si el paciente es mujer.
- RPDE: variable original.
- DFA: variable original.
- PPE: variable original.
- GNE: variable original.
- Jitter: variable resultado de la limpieza de valores extremos y análisis de componentes principales.
- Shimmer: variable resultado del análisis de componentes principales.

- MFCC1: variable resultado del análisis de componentes principales.
- HNR1: variable resultado del análisis de componentes principales.

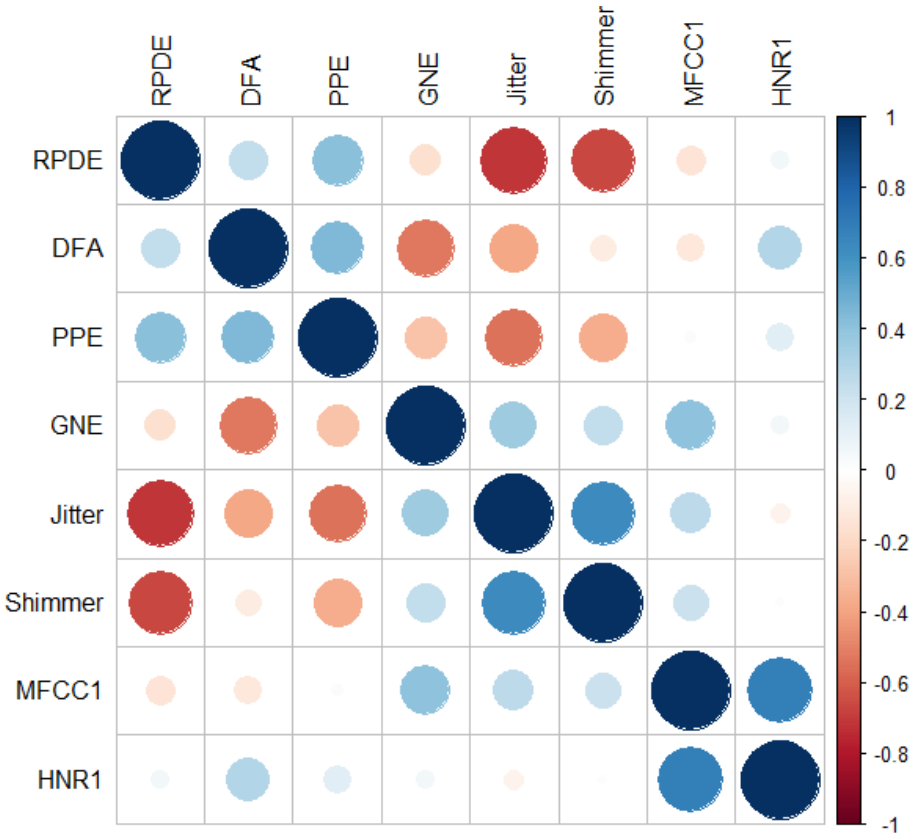


Figura 4.17: Correlación entre las variables resultantes.

Como se patentiza, en la figura 4.17 sobresale, por un lado, la correlación en la base de datos que nos ayudará con la modelación y, por otro lado, la existencia de una ligera correlación negativa entre la variable RPDE con las variables Jitter y Shimmer; mientras que en el resto de los casos la correlación fue mitigada en comparación con el conjunto de datos originales.

4.3. Modelación

Si se procura que, para cada individuo, se conozca la respuesta de la variable Status —denotada por 1, cuando el paciente sufre de Parkinson, y por 0, cuando el paciente se encuentra libre de Parkinson—, es necesario calcular la probabilidad del siguiente modo:

$$\mathbb{P}(Y_i = 1) = \pi_i$$
$$\Rightarrow \mathbb{P}(Y_i = 0) = 1 - \pi_i$$

Así pues, en esta sección se referirá el proceso que se llevó a cabo para encontrar un modelo apropiado a los datos de la presente investigación. Vale anticipar que se generaron dos tipos de modelos para ejemplificar la parte teórica descrita anteriormente:

- Modelos lineales generalizados.
- Modelos aditivos generalizados.

Pues bien, el análisis se llevará a cabo en varias fases, el primer ejemplo a tratar será un modelo de regresión binaria con las variables originales de la base de datos, y la variación (la función binaria) se aplicará en cada modelo; posteriormente, se ajustará un modelo de regresión aditivo y, finalmente, se compararán los resultados obtenidos con base en diferentes métricas.

Para terminar, es oportuno indicar que los datos se seccionarán en dos grupos, entrenamiento y evaluación, dicha segmentación se realizó mediante un muestreo aleatorio y se dividió la base de datos original con una proporción 80 % - 20 %.

4.3.1. Modelos lineales generalizados

El propósito central de esta sección es comparar el efecto obtenido al variar la función liga, utilizada para realizar la transformación en cada GLM. Dicho lo anterior, para el modelo se consideran las variables resultantes del análisis exploratorio; esto con el fin de expresar la condición de Parkinson con base en la relación que tienen con las variables y la asociación a la función liga.

	GLM 1	GLM 2	GLM 3
Función Liga:	Probit	Logit	Cloglog
(Intercepto)	11.27^{*}	19.18	13.46
	(5.72)	(10.41)	(7.16)
Gender	-1.06^{**}	-1.71^{**}	-1.19^{**}
	(0.33)	(0.58)	(0.38)
RPDE	-12.91^{***}	-20.91^{**}	-14.66^{**}
	(3.84)	(6.85)	(4.79)
DFA	-5.82^{*}	-10.09^{*}	-6.44^{*}
	(2.64)	(4.77)	(3.03)
PPE	1.29	2.28	1.67
	(0.78)	(1.38)	(0.91)
GNE	−3.86	−6.88	−5.91
	(5.56)	(10.10)	(6.97)
Jitter	1.58	2.57	1.63
	(0.81)	(1.39)	(0.91)
Shimmer	1.48	2.41	1.81
	(1.01)	(1.80)	(1.12)
MFCC1	-1.18^{***}	-2.05^{***}	-1.37^{***}
	(0.29)	(0.54)	(0.35)
HNR1	0.02^{**}	0.03^{*}	0.02^{*}
	(0.01)	(0.01)	(0.01)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tabla 4.1: Resumen estadístico de los modelos con su respectiva función liga.

Se probaron diferentes funciones liga con el mismo conjunto de datos, en la tabla 4.1 se muestra un resumen de los coeficientes calculados para cada una de las variables de acuerdo a cada transformación aplicada y en la tabla 4.2 un resumen de criterios para la evaluación de cada modelo.

	GLM 1	GLM 2	GLM 3
AIC	154.78	156.15	156.72
BIC	187.35	188.72	189.30
Log Likelihood	−67.39	−68.07	−68.36
Deviance	134.78	136.15	136.72
Num. obs.	192	192	192

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tabla 4.2: Resumen estadístico de los modelos con su respectiva función liga.

En resumen, se probaron tres funciones liga diferentes, cada una de ellas tiene una manera diferente de tranformación para el cálculo de las probabilidades correspondientes. A continuación se desglosan aquellas variables que fueron significativas en cada uno de los modelos:

- Cuando se ocupó la función probit como función liga, las variables más significativas para el modelo fueron RPDE y MFCC1, ambas afectando negativamente al pronóstico, es decir que por cada unidad que se modifique la probabilidad, y se moverá -12.91 veces RPDE.
- Cuando la función liga fue una función logit, se obtuvo un resultado similar con las variables, pero la variable más significativa es el MFCC1 y, en menor medida de significancia, las variables Gender y RPDE; en todos los casos las variables afectan negativamente a la probabilidad final.
- Finalmente, si se utiliza la función cloglog como función liga, se mantienen las variables anteriores, pero la relación entre ellas y la probabilidad calculada tiene una influencia menor que en el caso de la logit.

En el scatterplot de la figura 4.18 se presenta la distribución de la probabilidad estimada para cada variable, dicho sea de paso que: cuando se utiliza la función probit o la función logit es claro que la estimación por debajo de la línea es mejor pues se presentan más marcados los valores estimados con poca probabilidad, mientras que para la función cloglog ocurre un comportamiento contrario.

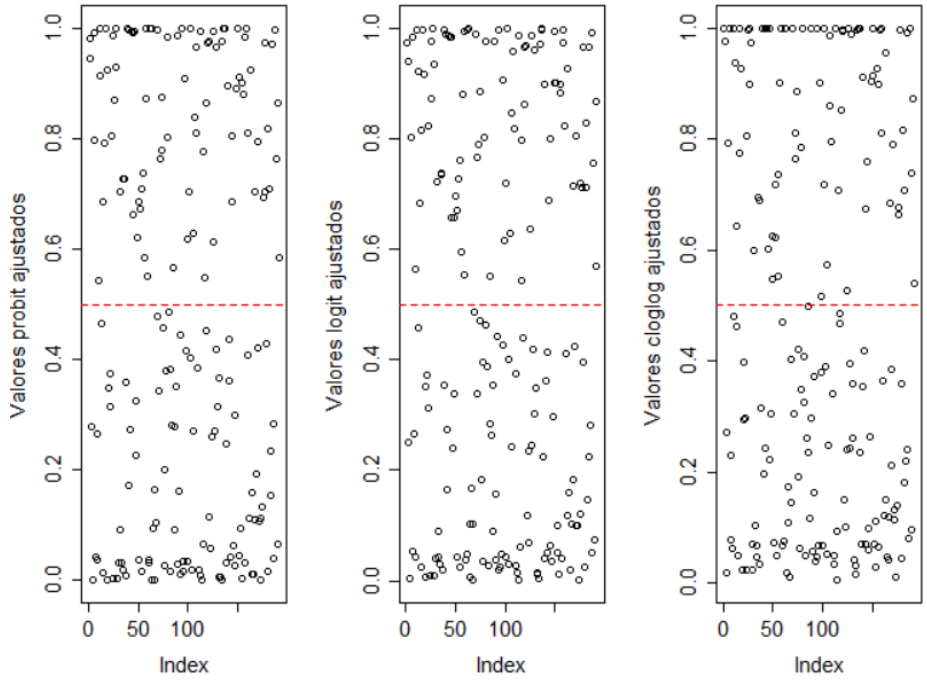


Figura 4.18: Presenta scatterplot de las probabilidades calculadas variando la función liga.

Ahora bien, cuando se elabora un modelo de regresión completo, muchas veces se pelea contra la mulcolinearidad en el modelo porque, cuando ésta se presenta, puede inflar la varianza de las variables del modelo y, además, se rompe el supuesto de independencia. Debido a la falta de existencia de una prueba estadística para determinar cuándo se enfrenta a la multicolinealidad, se utiliza el factor de inflación en la varianza (*VIF*, por sus siglas en inglés), el cual mide el impacto de la multicolinealidad para cada variable y se calcula por:

$$VIF_i = \frac{1}{1 - R_i}, \quad i = 1, \ldots, n$$

Como regla general, si el valor del VIF es mayor a 10, se traduce en la existencia de multicolinealidad en las covariables de la base de datos.

Entonces, al ocupar la función *vif* en R sobre el modelo lineal generalizado, sin aplicar las componentes principales ni el tratamiento de valores atípicos, esto es, al calcular sólo el VIF_i para cada una de las variables, se obtiene la tabla siguiente —donde en rojo se resaltan aquellas variables que obtuvieron un valor mayor a 10 (lo que refleja multicolinealidad en la base de datos original)—. Vale mencionar que, con base en esta información se intentó mejorar el modelo y, así, obtener una mejor aproximación al comportamiento de los datos.

Recording	Gender	Jitter_rel	Jitter_abs	Jitter_RAP	Jitter_PPQ	Shim_loc
1.43	4.39	93.42	21.02	67.21	76.24	1723.21
Shim_dB	Shim_APQ3	Shim_APQ5	Shi_APQ11	HNR05	HNR15	HNR25
329.89	641.85	75.18	84.88	51.28	381.15	342.21
HNR35	HNR38	RPDE	DFA	PPE	GNE	MFCC0
1200.11	975.93	9.69	12.47	3.32	4.21	6.32
MFCC1	MFCC2	MFCC3	MFCC4	MFCC5	MFCC6	MFCC7
5.18	5.65	4.86	5.07	4.72	5.37	5.81
MFCC8	MFCC9	MFCC10	MFCC11	MFCC12	Delta0	Delta1
9.64	6.81	6.73	6.52	5.93	4.74	4.13
Delta2	Delta3	Delta4	Delta5	Delta6	Delta7	Delta8
5.97	5.35	5.64	5.07	9.68	6.12	7.99
Delta9	Delta10	Delta11	Delta12			
5.30	6.59	4.75	6.19			

De hecho, al aplicar un modelo lineal generalizado se valida, primero, que se elimine la multicolinealidad derivada del filtrado de los componentes principales y del tratamiento que se le dio al valor atípico con el que se contaba, a continuación se muestran los VIF's de cada uno de los modelos:

Modelo	Gender	RPDE	DFA	PPE	GNE	Jitter	Shimmer	MFCC1	HNR1
GLM1	1.70	2.95	2.73	1.95	1.79	3.16	2.13	2.48	2.06
GLM2	1.77	3.02	3.01	2.03	1.79	3.13	2.16	2.46	2.04
GLM3	1.72	3.57	2.99	2.07	1.90	3.17	2.53	2.44	2.23

4.3.2. Modelos aditivos generalizados

En este apartado, a diferencia del anterior, el objetivo principal es conocer cómo afecta el uso de suavizamientos y de splines a la modelación o a la aproximación del modelo, también se busca plantear si al aumentar la complejidad de los mismos se dificulta la explicación de éste. Se utilizaron estimaciones a través de una aproximación aditiva, concentradas en tres tipos:

- **GAM 1**: GAM completo, donde se utlizan todas las variables que se han tratado y se aplica un suavizamiento a cada una de ellas.
- **GAM 2**: GAM completo, modificando el suavizamiento aplicado en las variables.
- **GAM 3**: GAM completo, aplicando un spline suavizado al modelo.

En el primer modelo aditivo, GAM 1, dado por:

$$Status = Gender + s(RPDE) + s(DFA) + s(PPE) + s(GNE)$$
$$+s(Jitter) + s(Shimmer) + s(MFCC1) + s(HNR1)$$

Al aplicarle el suavizamiento a cada una de las variables observamos que el coeficiente del suavizamiento es muy grande en su mayoría, presentado en la siguiente tabla, lo cual es

sinónimo de una mala aplicación del mismo, ya que cuando el coeficiente de suavizamiento es muy grande es señal de que la mejor aproximación es mediante una recta.

	Coef. de suavizamiento
s(RPDE)	4212.16
s(DFA)	17553.15
s(PPE)	110100.97
s(GNE)	275788.81
s(Jitter)	2.93
s(Shimmer)	0.00
s(MFCC1)	0.01
s(HNR1)	17601.57

Pues bien, en las gráficas de la figura 4.19 subsecuentes se muestra el efecto del modelo aditivo en cada una de las variables, cada pequeño círculo representa los residuos parciales de cada aproximación mientras que en azul se muestran bandas de confianza en cada variable.

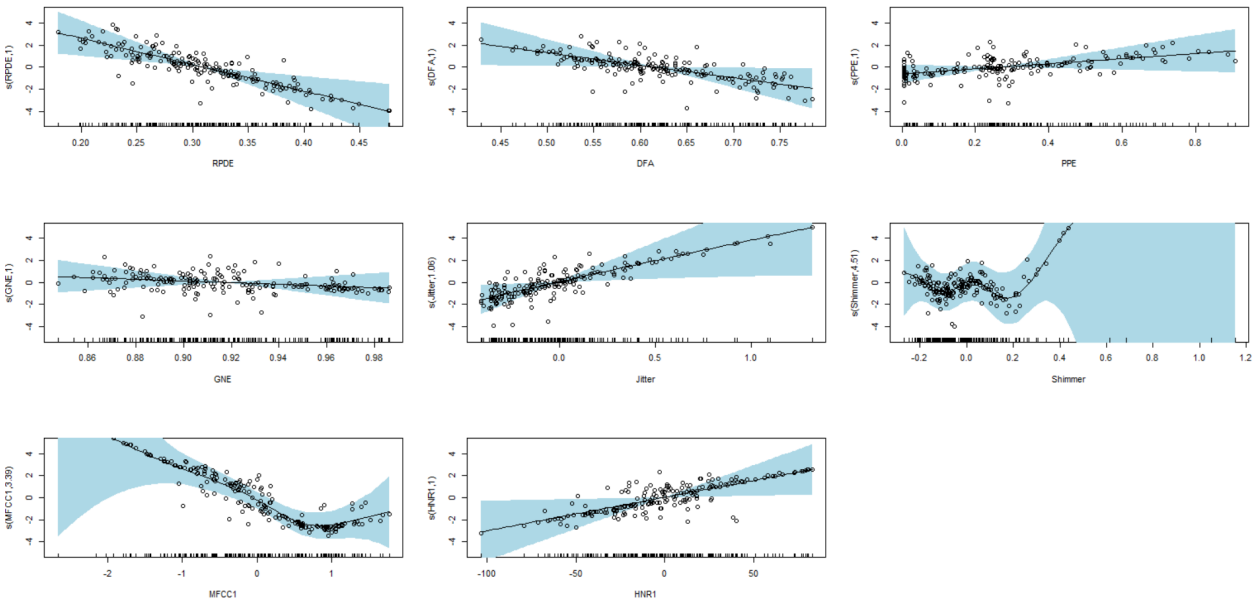


Figura 4.19: Gráfico del suavizamiento aplicado a cada variable.

Al generar un resumen del modelo, se obtiene el resultado siguiente:

```
Family: binomial
Link function: logit

Formula:
Status ~ Gender + s(RPDE) + s(DFA) + s(PPE) + s(GNE) + s(Jitter) +
    s(Shimmer) + s(MFCC1) + s(HNR1)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.2269     0.8681   1.413  0.15756
Gender1      -1.8452     0.6493  -2.842  0.00448 **
---
```

```
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Approximate significance of smooth terms:
             edf Ref.df Chi.sq  p-value
s(RPDE)    1.000  1.000 10.626 0.001116 **
s(DFA)     1.000  1.000  4.639 0.031260 *
s(PPE)     1.000  1.000  2.241 0.134381
s(GNE)     1.000  1.000  0.523 0.469538
s(Jitter)  1.063  1.123  5.639 0.021540 *
s(Shimmer) 4.507  5.318  5.798 0.542352
s(MFCC1)   3.387  4.256 21.189 0.000434 ***
s(HNR1)    1.000  1.000  4.939 0.026253 *
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

R-sq.(adj) =  0.569   Deviance explained = 56.4%
UBRE = -0.23018  Scale est. = 1         n = 192
```

A colación de lo preliminar, la columna edf (grados efectivos de libertad, por sus siglas en inglés) se interpreta como el número de funciones base necesarias para el suavizamiento, confirmando de este modo que las únicas variables para aplicar el suavizamiento son Shimmer y MFCC1 con 4 y 3 funciones base, mientras que el resto sería conveniente aproximarlas con una función lineal.

De ello resulta la generación del siguiente modelo, GAM 2, donde se conservan únicamente tres suavizamientos; tal modelo es expresado por:

$$\eta(Status) = Gender + RPDE + DFA + PPE + GNE + s(Jitter) + s(Shimmer) + s(MFCC1) + HNR1$$

Puesto que no se modificó el número de funciones base en los suavizamientos, la gráfica es la misma en las variables y la aproximación; en este caso, al hacer una síntesis del modelo se observa que se agrega el intercepto como parámetro significativo y se reduce el valor del p-value en las variables que ya eran significativas, como se muestra a continuación:

```
Family: binomial
Link function: logit

Formula:
Status ~ Gender + RPDE + DFA + PPE + GNE + s(Jitter) + s(Shimmer) +
    s(MFCC1) + HNR1

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  22.01666   10.86226   2.027  0.04267 *
Gender1      -1.84527    0.64926  -2.842  0.00448 **
RPDE        -23.98211    7.35636  -3.260  0.00111 **
DFA         -11.41436    5.29909  -2.154  0.03124 *
PPE           2.30642    1.54099   1.497  0.13447
GNE          -7.53913   10.42694  -0.723  0.46965
HNR1          0.03045    0.01370   2.223  0.02624 *
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Approximate significance of smooth terms:
             edf Ref.df Chi.sq  p-value
s(Jitter)  1.066  1.127  5.640 0.021687 *
s(Shimmer) 4.506  5.317  5.794 0.542610
s(MFCC1)   3.387  4.257 21.187 0.000435 ***
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

R-sq.(adj) =  0.569   Deviance explained = 56.4%
UBRE = -0.23018  Scale est. = 1         n = 192
```

En cuanto al valor del coeficiente del suavizamiento, todos toman valores pequeños, por tanto, en cuestión del Jitter, no afecta que los grados efectivos de libertad sean iguales a 1.

	s(Jitter)	s(Shimmer)	s(MFCC1)
GAM 2	2.82	0.00	0.01

A propósito del GAM3, se ejemplifica el uso de splines, y, a diferencia de los anteriores, el modelo está dado por:

$$Status = Gender + RPDE + DFA + PPE + GNE + s(Jitter) + s(Shimmer) + s(MFCC1) + HNR1$$

donde cada suavizamiento es uno penalizado (p-spline). Después de efectuar un prontuario del modelo obtenido, se descubre lo siguiente:

```
Family: binomial
Link function: logit

Formula:
Status ~ Gender + RPDE + DFA + PPE + GNE + s(Jitter, bs = "ps") +
    s(Shimmer, bs = "ps") + s(MFCC1, bs = "ps") +
    HNR1

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  22.47054   11.16628   2.012 0.044182 *
Gender1      -1.72944    0.63203  -2.736 0.006213 **
RPDE        -25.73655    7.37899  -3.488 0.000487 ***
DFA         -11.37027    5.27028  -2.157 0.030972 *
PPE           2.20613    1.47527   1.495 0.134810
GNE          -7.90430   10.68101  -0.740 0.459280
HNR1          0.02845    0.01314   2.165 0.030362 *
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Approximate significance of smooth terms:
             edf Ref.df Chi.sq  p-value
s(Jitter)  1.664  2.065  5.607 0.064253 .
s(Shimmer) 1.997  2.478  2.611 0.354610
s(MFCC1)   3.726  4.414 20.887 0.000737 ***
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

R-sq.(adj) =  0.552   Deviance explained = 54.4%
UBRE = -0.21827  Scale est. = 1         n = 192
```

Comparando con el modelo anterior —por añadir un suavizamiento más complejo y esto, a su vez, genera un spline penalizado por cada variable suavizada—, se observa que no necesariamente al aumentar la complejidad del modelo se obtiene una mejor aproximación pues, al realizar el coeficiente del suavizamiento, ésta aumenta. Por su parte, en la variable Jitter el coeficiente aumenta, misma que podría ser mejor aproximada por una recta.

	s(Jitter)	s(Shimmer)	s(MFCC1)
GAM 2	2.82	0.00	0.01
GAM 3	49.61	6.33	0.98

Tras graficar la aproximación aplicando la penalización, se observa que (ver figura 4.20), en cuanto a la variable MFCC1 se refiere, es donde mejor se comporta; mientras que en la

variable Jitter —en la cual se presenta el mayor aumento del coeficiente de suavizamiento— es aproximada por una curva casi logarítmica.

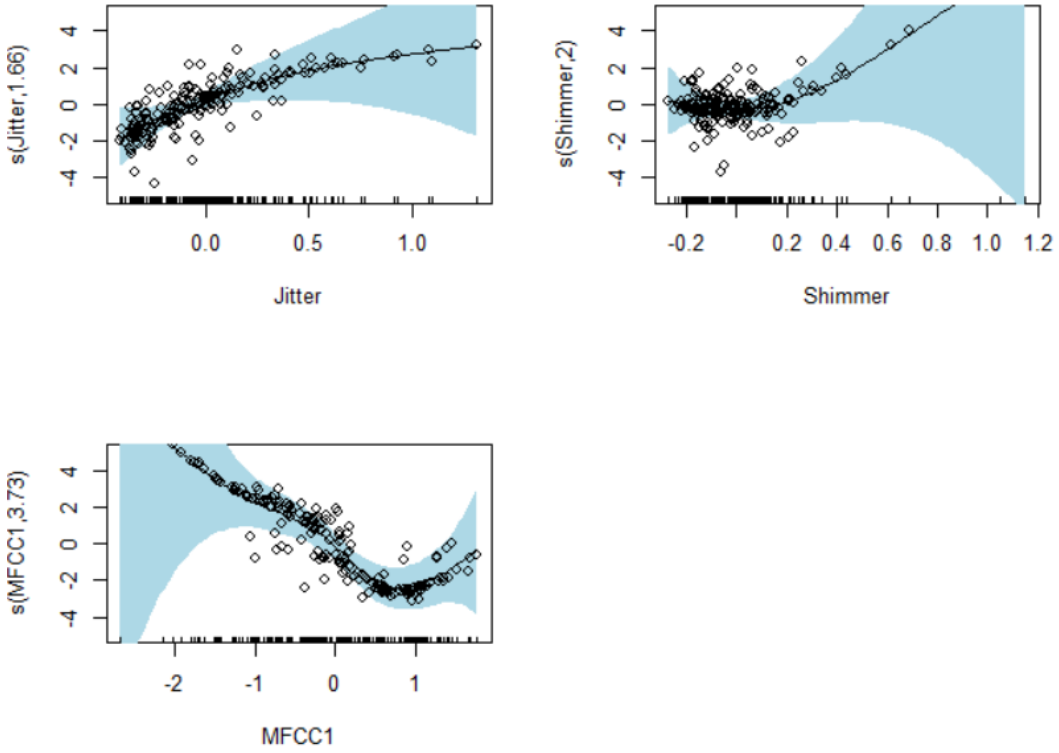


Figura 4.20: Gráfico del suavizamiento aplicado a cada variable.

En el caso del tercer modelo, GAM3, se encuentra menor acumulación de estimaciones cercanas al 0.5 (ver figura 4.21), lo cual implica que la tasa de mala clasificación tiene una banda más alta, alrededor del 0.5, como para mostrar ese cambio.

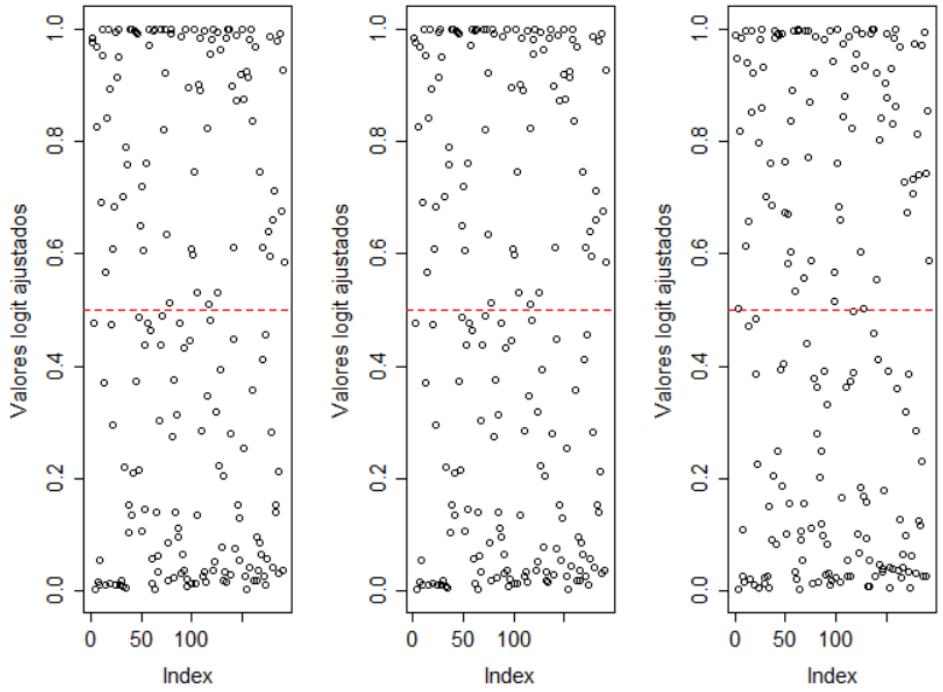
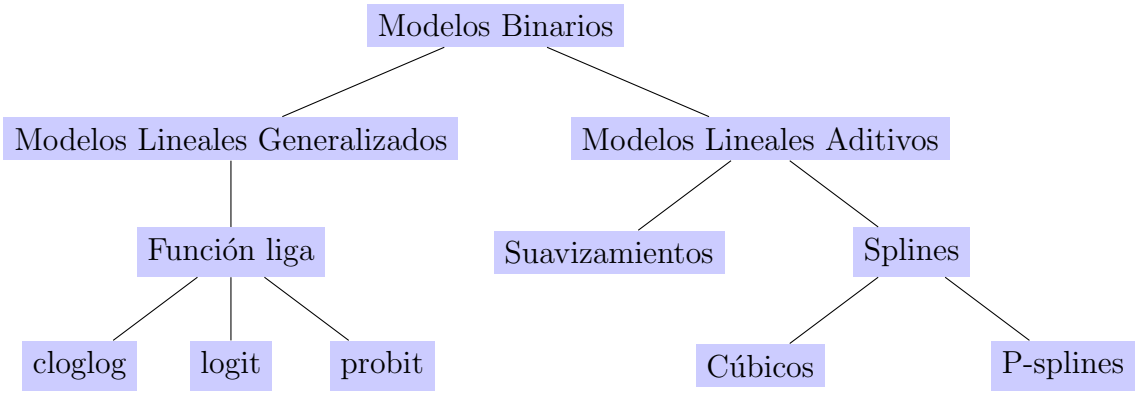


Figura 4.21: Presenta scatterplot de las probabilidades calculadas variando el tipo de suavizamiento. Iz. GAM1, Centro GAM2, Der. GAM3.

4.3.3. Evaluación del modelo

Recapitulando, los modelos realizados se pueden clasificar de la siguiente manera:



Ahora bien, para validar cuál modelo es el ganador, pueden tomarse en cuenta diferentes métricas, a continuación se listan algunas de las que se usan en esta sección:

- AIC, como medida basada en la verosimilitud.
- Curva ROC, donde además analizamos de manera individual la sensibilidad y la especificidad.

En la siguiente tabla se muestra el resumen de un modelo lineal generalizado utilizando la base de datos completo, para ello se empleó la función liga predeterminada, logit, y no

se realizó ninguna modificación como tratamiento de outliers ni se redimensionó mediante componentes principales; este modelo sirve de referencia para medir el desempeño de los modelos calibrados.

Entonces, si se considera el criterio de información de Akaike (AIC) para los modelos lineales generalizados, el mejor modelo es el modelo GLM 1 —el cual está asociado a la función liga probit— porque es el modelo con el menor valor.

Por otro lado —si se atienden únicamente los modelos aditivos y si se prescinde de criterios como aplicar el suavizamiento a todas las variables porque disminuye la complejidad del modelo—, hay una paridad entre GAM 1 y GAM 2; de lo contrario, el ganador sería GAM 2, a saber, cuando se aumentó la complejidad de los suavizamientos, se disminuyeron los grados de libertad del modelo final, pero el AIC se aumentó en tres unidades, por lo que aumentar la complejidad no propició una mejor aproximación.

	df	AIC
GLM 1	11.00	156.32
GLM 2	11.00	157.64
GLM 3	11.00	158.58
GAM 1	15.96	147.81
GAM 2	15.96	147.81
GAM 3	14.39	150.09

En el mismo orden de ideas, otra forma de conocer el desempeño de nuestro modelo es mediante la curva ROC pues gracias a ésta se determina el punto ideal para el corte de la probabilidad y se minimza el error estimado. A continuación se grafica la *Sensibilidad* contra $1 - Especificidad$, donde se pondrá atención sobre todo en el Área Bajo la Curva (AUC, por sus siglas en inglés).

Pues bien, el punto de corte de los primeros dos modelos GLM es muy cercano a 0.5 (ver figura 4.22), aunque el AUC es muy cercano a 0.9 cuando se cumple este punto de corte; es decir, estableciendo 0, cuando $p < 0.5$ y 1 en otro caso, la estimación generada es muy buena.

En cuanto al último GLM (ver figura 4.22), el punto de corte de la probabilidad es menor, cercano a 0.46, obteniendo así la Sensibilidad y la Especificidad ideales; esto es que, por lo que se puede inferir, predice mejor a quienes no presentan la enfermedad que los pacientes con Parkinson o unos en el modelo. En síntesis, de los modelos aditivos: el GAM 1 presenta un aumento en la curva de 0.8 y éste indica el valor ideal, en tanto que el AUC es cercano al 0.9, por lo que —cuando se toma un punto de corte menor a 0.88— no se nota diferencia alguna entre la cantidad de Falsos Positivos y Verdaderos Negativos estimados.

Para concluir, en los últimos dos modelos aditivos calculados (ver figura ROC) —a pesar de la variación en el tipo de suavizamiento—, no se gana una ventaja considerable en la estimación. Sirva de observación adicional que, en el caso de los modelos aditivos, es mejor la estimación que un modelo lineal pues, cuando se presenta una probabilidad mayor al 0.8, es seguro que se presente la condición de Parkinson en el paciente, a diferencia de los modelos lineales que, cuando la probabilidad es cercana a 0.5, son azarosos respecto a la estimación de contar con Parkinson o no.

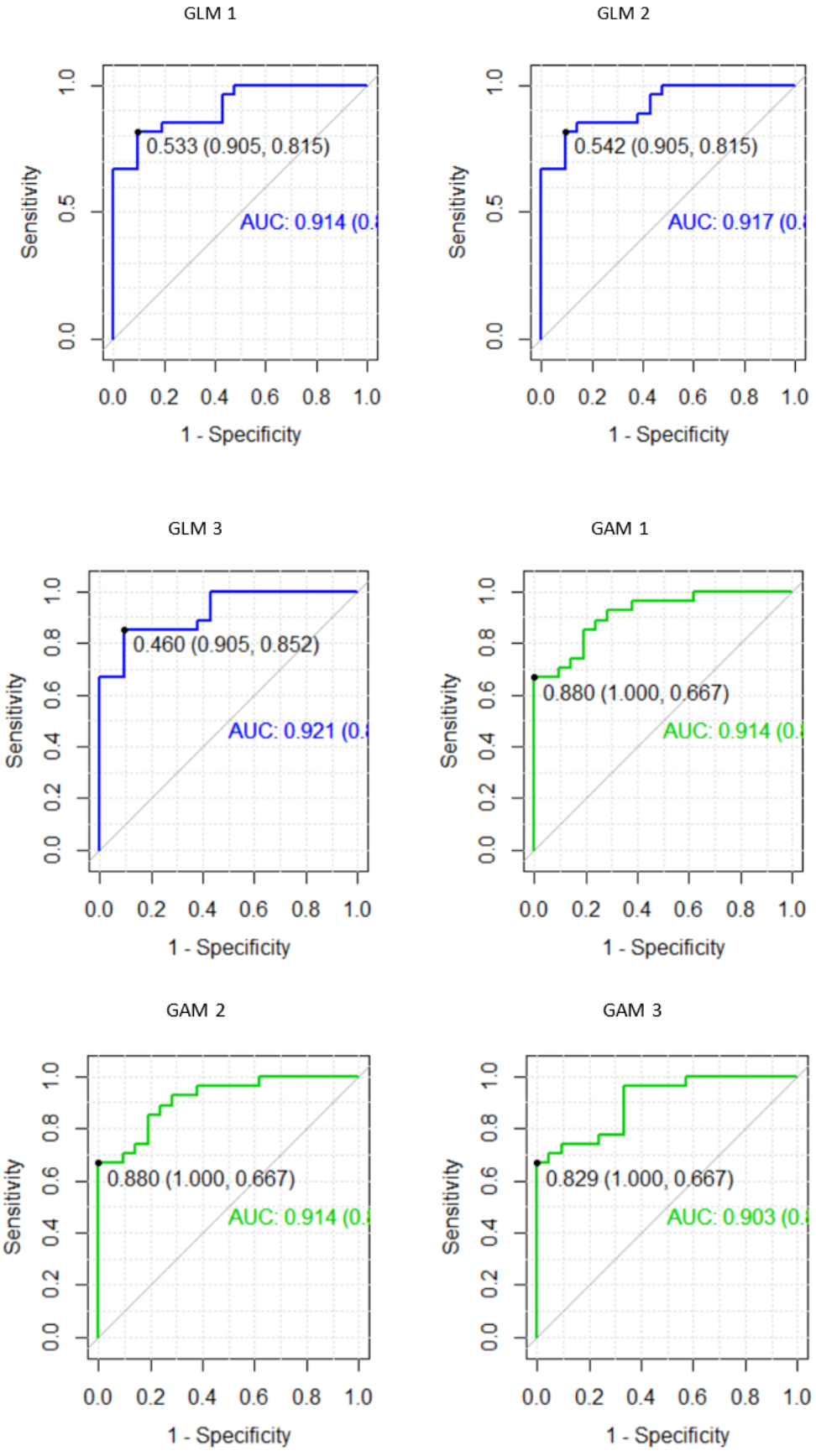


Figura 4.22: Curva ROC para cada modelo.

Capítulo 5

Conclusiones

El desarrollo de este proyecto se centró en el estudio de los modelos aditivos generalizados y se analizó el desempeño de éstos frente a un modelo lineal. Como se constató, existen diversas diferencias entre un modelo lineal y un modelo aditivo, la más sustancial es que —incluso en un escenario con poca información o utilizando algún método de selección de variables— un modelo lineal no tiene mucha ventaja respecto al ajuste de un modelo aditivo; además, los modelos aditivos permiten explicar relaciones que no siguen un comportamiento lineal, lo cual permite una mayor flexibilidad y libertad al elegir el tipo de covariables a modelar.

Mientras que, una ventaja de ambos modelos generalizados, tanto lineales como aditivos, es la facilidad de modelar diferentes variables respuesta, con el simple hecho de pertenecer a la familia exponencial, lo cual abre un abanico de posiblidades a nuevas aproximaciones y soluciones a diferentes problemas prácticos.

Sin embargo, la principal debilidad de un GAM se encuentra en la interpretabilidad del modelo, dado que aumenta la complejidad del mismo. En otras palabras, cuando se tiene un modelo lineal, el coeficiente calculado es directamente proporcional a la cantidad de información que aporta dicha variable; en el caso del GAM se calcula una función determinada por el coeficiente del suavizamiento por lo que se gana en ajuste, pero se pierde la interpretación directa del mismo.

De acuerdo con los resultados obtenidos, detallados en el capítulo 4, se pueden extractar del siguiente modo:

- Se obtiene un mejor ajuste y una mejor estimación como beneficio de obtener un modelo aditivo como modelo ganador.
- Se ganan grados de libertad, por lo que aumenta la complejidad del modelo.
- Dado que se cuenta con un número limitado de observaciones o repeticiones del experimento, no es viable probar con más tipos de suavizamiento ni siquiera segmentando por alguna variable categórica, como género.

Cabe señalar que, si bien se expusieron máxime las diferencias en cuanto al empleo de estas dos técnicas, las utilidades de éstas no se acotan a este universo, existen diferentes enfoques por los cuales darle uso a un modelo aditivo y diferentes formas de ajustar los suavizamientos, propiciando así un enfoque de modelos mixtos (GAMM) —el cual comprende los modelos de regresión, simple y de polinomios, modelos discretos, modelos binarios,

modelos multinomiales, entre otros—. De hecho, en estos modelos mixtos también es posible encontrar, por una parte, su análogo para el caso en que la variable respuesta es un vector: modelos aditivos generalizados vectoriales (VGAM) y, por otra, su versión lineal: modelos lineales generalizados vectoriales (VGLM) y, de igual manera, no se acotan a variables respuestas continuas.

También han sido aplicados en modelos de series de tiempo y pueden ser competitivos al nivel de los modelos no paramétricos, como un bosque aleatorio o ganar interpretabilidad respecto a una red neuronal. Igualmente, en la literatura es posible encontrar aplicaciones en series de tiempo, siendo un GAM el algoritmo usado iterativamente para el ajuste y, así, reducir el error en la predicción; todo esto al identificar y resumir múltiples funciones, lo que produce un mejor ajuste en la tendencia de la serie.

Bibliografía

[Bazaraa *et al.*, 2004] Bazaraa, M., Jarvis, J., y Sherali, H. (2004). *Programación lineal y flujo en redes*. Limusa, México, DF, 2da edición.

[Berk, 2008] Berk, R. A. (2008). *Statistical Learning from a Regression Perspective*. Springer.

[Bhat y Kumar, 2012] Bhat, H. S. y Kumar, N. (2012). On the derivation of the bayesian information criterion. *IMMAGIC Sitio web*. `https://www.immagic.com/eLibrary/ARCHIVES/GENERAL/WIKIPEDI/W120607B.pdf`.

[Binder y Tutz, 2008] Binder, H. y Tutz, G. (2008). A comparison of methods for the fitting of generalized additive models. *Stat Comput*, 18:87–99.

[del Carmen García *et al.*, 2014] del Carmen García, M., Castellana, N., Rapelli, C., Koegel, L., y Catalano, M. (2014). Criterios de información y predictivos para la selección de un modelo lineal mixto. *SaberES*, N 6:61 – 74.

[Dias *et al.*, 2016] Dias, A., Limongi, J., Barbosa, E., y WT, H. (2016). Voice telerehabilitation in Parkinson's disease. telerreabilitação vocal na doença de Parkinson. *Codas*.

[Hastie y Tibshirani, 1990] Hastie, T. y Tibshirani, R. (1990). *Generalizaed Additive Models*. Chapman & Hall/CRC, 1st edición.

[Honaker *et al.*, 2012] Honaker, J., King, G., y Blackwell, M. (2012). *AMELIA II: A Program for Missing Data, 2021*. `https://r.iq.harvard.edu/docs/amelia/amelia.pdf`.

[Pérez, 2016] Pérez, C. J. (2016). Parkinson dataset with replicated acoustic features data set. `https://archive.ics.uci.edu/ml/datasets/Parkinson+Dataset+with+replicated+acoustic+features+`.

[Teixeira *et al.*, 2013] Teixeira, J., Oliveira, C., y Lopes, C. (2013). Vocal acoustic analysis - jitter, shimmer and hnr parameters. *Procedia Technology*, 9.

[Yee, 2015] Yee, T. W. (2015). *Vector Generalized Linear and Additive Models With an Implementation in R*. Springer.