Grado en Matemáticas

TRABAJO FIN DE ESTUDIOS

Modelos aditivos generalizados en análisis del cambio climático

Tutor: José Luis Pino Mejías

Alumno: Francisco José Lozano Ruiz

Sevilla, Octubre de 2023

Índice general

Prólogo III
Resumen V
Abstract VI
Índice de Figuras VII
Índice de Tablas IX

1. Introducción **1**

2. Modelos lineales generalizados **3**
2.1. Introducción 3
2.2. Modelos lineales 4
2.3. Modelos lineales generalizados 6
2.3.1. Familia de distribuciones exponenciales 7
2.3.2. Ajuste de los modelos lineales generalizados 9
2.3.3. Funciones de enlace canónicas 11
2.3.4. Residuos 11

3. Modelos aditivos generalizados **13**
3.1. Introducción 13
3.2. Suavizado univariante 14
3.2.1. Bases de funciones 14
3.2.2. Control del suavizado 16
3.2.3. Elección del parámetro de suavizado 17
3.3. Modelos aditivos 18
3.4. Smoothers 20
3.4.1. Splines cúbicos 21
3.4.2. Smoothers unidimensionales 23
3.5. Modelos aditivos generalizados 24
3.5.1. Ajuste del modelo 25

4. Título del Capítulo **27**

4.1. Primera sección 27

A. Apéndice: Título del Apéndice **29**

A.1. Primera sección 29

B. Apéndice: Título del Apéndice **31**

B.1. Primera sección 31

Bibliografía **33**

Prólogo

Escrito colocado al comienzo de una obra en el que se hacen comentarios sobre la obra o su autor, o se introduce en su lectura; a menudo está realizado por una persona distinta del autor.

También se podrían incluir aquí los agradecimientos.

Resumen

Resumen...

Abstract

Abstract...

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El modelado estadístico es una herramienta fundamental para la investigación científica y el análisis de datos, su principal propósito es el de aproximar la realidad a partir de la implementación de modelos matemáticos que tienen en cuenta la incertidumbre. Estos tipos de modelos son capaces de abarcar distintos problemas como pueden ser: la descripción de relaciones entre variables, la predicción de nuevos datos o la comprobación de hipótesis.

Hoy en día existen muchos métodos y técnicas para proceder a resolver los problemas antes mencionados, pero en este trabajo nos centraremos en el desarrollo de los Modelos Aditivos Generalizados (MAG). Sin embargo, hasta llegar a ellos, pasaremos por la descripción de los Modelos Lineales, los Modelos Lineales Generalizados (MLG) y los Modelos Aditivos. Esto se debe a que los MAG no son más que una extensión de los anteriores, así que haremos un transcurso desde modelos simples hasta un Modelo Aditivo Generalizado completo.

En el primer capítulo hablaremos de los Modelos Lineales y los Modelos Lineales Generalizados. El concepto de regresión lineal surgió a partir de la necesidad de estudiar la relación entre unas variables, de las cuales se conocen ciertos datos, mediante formalizaciones matemáticas. En concreto lo introdujo Francis Galton y luego fue desarrollado por el estadístico y matemático Karl Pearson a finales del siglo XIX. Sin embargo, ya en el 1805 Legendre proponía la primera forma del método de mínimos cuadrados, por lo que estamos hablando de técnicas que ya llevan más de dos siglos entre nosotros. A pesar de ello, no se desarrollan las nociones de los MLG hasta el 1970, estos modelos relajan las hipótesis que deben asumir los Modelos Lineales y permiten un primer acercamiento a que los modelos tengan un grado de no linealidad.

Comenzaremos el siguiente capítulo introduciendo los Modelos Aditivos y varios resultados básicos sobre la estimación de los MAG, a partir de ellos ya nos podremos adentrar más en profundidad en los resultados necesarios para sus futuras aplicaciones. En particular, añadiremos una sección que hable de las funciones de suavizado (*smoothers*) y los tensores (*tensors*) que aplicaremos luego en la práctica, además de presentar también conceptos sobre la comparación de estos tipos de modelos y sobre el contraste de hipótesis.

Los Modelos Aditivos Generalizados mejoran la metodología de los MLG incorporando la flexibilidad que aporta la regresión no paramétrica y mantienen la interpretabilidad de los datos del análisis de regresión con múltiples variables predictoras pues se modelan

como una suma de términos 'suaves'. Actualmente, estos modelos son un punto de partida magnífico para el modelado de un problema, un GAM bien ajustado debería funcionar de manera conveniente, incluso comparado con métodos de *boosting* o de *deep learning*. Además, el MAG tiene una base para una mejor interpretabilidad y métricas de incertidumbre más sencillas, por lo que en muchos casos del análisis de datos, los Modelos Aditivos Generalizados son una buena opción.

Capítulo 2

Modelos lineales generalizados

2.1. Introducción

Los modelos estadísticos pretenden explicar la relación entre dos o mas variables, en particular, tratan de describir el comportamiento de una variable respuesta (o dependiente), que se suele denotar por Y, mediante la información que otorgan las variables predictoras (o independientes), que se suelen denotar como $X_1, ..., X_p$.

Como se indica en James et al. [2014], la forma más general de expresar matemáticamente los modelos estadísticos es la siguiente:

$$Y = f(X_1, .., X_p) + \epsilon \tag{2.1}$$

Donde f es una función desconocida cuyo propósito es el de representar de la mejor[1] manera posible la relación entre las variables $X_1, ..., X_p$ e Y; y ϵ es un error aleatorio independiente de las variables predictoras que deberá cumplir ciertas condiciones según el tipo de modelo que estemos tratando.

Este trabajo tiene como finalidad el definir un conjunto de estrategias y técnicas que proporcionan un ajuste óptimo de la función f, es decir, que asemeje lo mejor posible la relación de las variables al fenómeno que se esté estudiando. Además, se incluye una posterior aplicación práctica de dichos resultados en el ámbito del cambio climático.

En lo que a este capítulo respecta, partiremos definiendo los modelos lineales de una forma breve y más general, ya que se ve de manera más extensa en varias asignaturas durante el grado. Tras ello, daremos varios resultados básicos para el entendimiento y desarrollo de los Modelos Lineales Generalizados.

[1] Se supone que se los nodos están espaciados de manera uniforme, pues en el caso de no que no lo estuvieran habría que añadir pesos a la suma.

2.2. Modelos lineales

El modelo lineal ocupa un lugar clave en el manual de herramientas de todo estadístico aplicado. Esto se debe a su simple estructura, a la fácil interpretación de sus resultados y al sencillo desarrolo de la teoría de mínimos cuadrados. Sin embargo, a la hora de dar su definición se deben tener en cuenta ciertas restricciones que deben cumplir las variables y los errores del modelo. Estas condiciones hacen que el modelo no sea capaz de adaptarse bien a todos los fenómenos que uno propone describir con él pero, a cambio, otorga esa sencillez de visualización antes mencionada. Daremos a continuación una serie de definiciones y resultados basados en Wood [2017].

Definición 2.2.1 (Modelo Lineal)**.** Sean $X_1, \ldots, X_p$ un conjunto de p vectores aleatorios de n componentes (con $p \leqslant n$) e $Y = (Y_1, \ldots, Y_n)^T$ un vector aleatorio de n componentes tal que $E[Y] = \mu$. Entonces, se entiende por modelo lineal (multivariante) aquel que determina la relación entre los vectores aleatorios mediante una combinación lineal de parámetros de la siguiente forma:

$$\mu = X\beta$$

$$Y = \mu + \epsilon$$

O vectorialmente como:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{(1,1)} & \cdots & x_{(1,j)} & \cdots & x_{(1,p)} \\ \cdots & \ddots & \vdots & \ddots & \cdots \\ x_{(i,1)} & \cdots & x_{(i,j)} & \cdots & x_{(i,p)} \\ \cdots & \ddots & \vdots & \ddots & \cdots \\ x_{(n,1)} & \cdots & x_{(n,j)} & \cdots & x_{(n,p)} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_i \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Donde:

- β es un vector de parámetros β_i a determinar que reflejan la magnitud del efecto lineal (constante) de los incrementos unitarios en las variables explicativas X_i sobre la variable explicada Y.
- ϵ es un vector aleatorio tal que los ϵ_i son variables aleatorias independientes e identicamente distribuidas por una distribución normal de esperanza nula y varianza σ^2 ($\epsilon_i \backsim N(0, \sigma^2)$). Representa el término de error del modelo y corresponde con $\epsilon = Y - E[Y]$.

Observación 2.2.1. Gracias a lo notado en el párrafo anterior podemos ver que:
$Y \backsim N(\mu, \sigma^2 I_n)$ ya que:

$$E[Y] = E[X\beta + \epsilon] = X\beta + E[\epsilon] = X\beta = \mu$$

$$Cov(Y) = Cov(X\beta + \epsilon) = Cov(\epsilon) = \sigma^2 I_n$$

Veamos ahora cómo se pueden obtener los valores de los parámetros β.

Definición 2.2.2 (Estimador por mínimos cuadrados)**.** Elegiremos los valores de β que minimicen la suma de cuadrados:

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \mu_i)^2 = \sum_{i=1}^{n} (Y_i - (X\beta)_i)^2 \tag{2.2}$$

Es decir:

$$\hat{\beta} = arg_{\beta \in \mathbb{R}^p} min ||Y - X\beta||^2$$

Donde $||\cdot||$ denota la norma euclídea.

Para encontrar tal mínimo se razona derivando respecto de cada β_i y luego igualando a 0. De este modo los parámetros β_i vienen dados por la solución del siguiente sistema:

$$\begin{cases} \frac{\partial S}{\partial \beta_1} = 0 \\ \cdots \\ \frac{\partial S}{\partial \beta_p} = 0 \end{cases}$$

Desarrollando este sistema de escuaciones llegamos a que es equivalente a $X^T X\hat{\beta} = X^T Y$, por lo que el estimador por mínimos cuadrados del vector de parámetros β viene dado por:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2.3}$$

Observación 2.2.2. Notemos que esta expresión tiene sentido pues el producto $X^T X$ resulta en una matriz cuadrada de orden p y rango máximo, por lo que el sistema antes dado tiene solución única.

Proposición 2.2.1 (Distribución del estimador $\hat{\beta}$)**.** El estimador por mínimos cuadrados del vector de parámetros β, $\hat{\beta}$, sigue una distribución del tipo normal p-variante de esperanza β y matriz de covarianzas $V_{\hat{\beta}} = \sigma^2 (X^T X)^{-1}$. Es decir: $\hat{\beta} \backsim N_p(\beta, \sigma^2 (X^T X)^{-1})$

Demostración. Partiremos viendo que el estimador $\hat{\beta}$ es insesgado:

$$E[\hat{\beta}] = E[(X^T X)^{-1} X^T Y] = (X^T X)^{-1} X^T E[Y] = (X^T X)^{-1} X^T X\beta = \beta$$

Por otro lado, calculemos la matriz de covarianzas de $\hat{\beta}$, para ello primero debemos notar que como los errores aleatorios ϵ_i son independientes e identicamentes distribuidos con esperanza nula y varianza σ^2, $\forall i \neq j$:

$$E[\epsilon_i \epsilon_j] = E[\epsilon_i] + E[\epsilon_j] + Cov(\epsilon_i, \epsilon_j) = 0$$

y, por tanto:

$$E[\epsilon\epsilon^T] = E\begin{bmatrix} \epsilon_1^2 & \epsilon_1\epsilon_2 & \cdots & \epsilon_1\epsilon_n \\ \epsilon_1\epsilon_2 & \epsilon_2^2 & \cdots & \epsilon_2\epsilon_n \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_1\epsilon_n & \cdots & \cdots & \epsilon_n^2 \end{bmatrix} = \begin{pmatrix} E[\epsilon_1^2] & E[\epsilon_1\epsilon_2] & \cdots & E[\epsilon_1\epsilon_n] \\ E[\epsilon_1\epsilon_2] & E[\epsilon_2^2] & \cdots & E[\epsilon_2\epsilon_n] \\ \vdots & \vdots & \ddots & \vdots \\ E[\epsilon_1\epsilon_n] & \cdots & \cdots & E[\epsilon_n^2] \end{pmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma^2 \end{bmatrix}$$

Luego, utilizando que por 2.3 se tiene que $\hat{\beta} = \beta + (X^T X)^{-1} X\epsilon$, obtenemos que la matriz de covarianzas es:

$$\begin{aligned} Cov(\hat{\beta}) &= E[(\hat{\beta} - E[\hat{\beta}])(\hat{\beta} - E[\hat{\beta}])^T] = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = \\ &= E[((X^TX)^{-1}X^T\epsilon)((X^TX)^{-1}X^T\epsilon)^T] = E[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-T}] = \\ &= (X^TX)^{-1}X^TE[\epsilon\epsilon^T]X(X^TX)^{-T} = (X^TX)^{-1}X^T\sigma^2I_nX(X^TX)^{-T} = \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-T} = \sigma^2(X^TX)^{-1} \end{aligned}$$

Acabamos la prueba recordando que $\hat{\beta}$ no es más que una combinación lineal de variables aleatorias normales, concretamente de las Y_i, para decir que en efecto sigue una distribución normal p-variante como indica el enunciado. ■

Observación 2.2.3. Ahora bien, generalmente el valor de σ^2 es desconocido así que también sería preciso dar una estimación del mismo para que así los resultados anteriores fueran de alguna utilidad.

Definición 2.2.3 (Estimador de σ^2)**.** La varianza σ^2 admite un estimador insesgado que se basa en la suma de cuadrados:

$$\hat{\sigma}^2 = \frac{S}{n-p} = \frac{\sum_{i=1}^{n}(Y_i - (X\beta)_i)^2}{n-p}$$

Además se tiene que $\hat{\sigma}^2$ sigue una distribución:

$$\hat{\sigma}^2 \backsim \frac{\sigma^2}{n-p}\chi^2_{n-p}$$

Observación 2.2.4. No daremos la obtención de este estimador pero sí indicaremos que su distribución se obtiene directamente de que como S es la suma de normales $N(0,\sigma^2)$ y $\frac{\epsilon_i}{\sigma} \backsim N(0,1)$, entonces: $\frac{1}{\sigma^2}S \backsim \chi^2_{n-p}$. Finalmente, multiplicando y dividiento esta expresión por (n-p) y sustituyendo la definición del estimador $\hat{\sigma}^2$ obtenemos el resultado.

2.3. Modelos lineales generalizados

Nos adentramos ya con esta sección en la primera extensión de los modelos lineales de las que se tratan durante el trabajo. Los Modelos Lineales Generalizados (MLG) fueron originalmente formulados por John Nelder y Robert Wedderburn (1972), quienes tenían como propósito unificar varios modelos estadísticos como la regresión lineal, la logística y la de Poisson en un mismo modelo. Este tipo de modelo relaja algunas de las hipótesis que asumían los modelos lineales, como que los errores ya no deben seguir ninguna distribución específica, y además añade nuevos elementos como la función de enlace, la cual interviene en la relación entre los valores esperados y la forma lineal del modelo. También se deberá tener en cuenta una nueva hipótesis distribucional, las variables de respuestas seguirán distribuciones de tipo exponencial. Más tarde introduciremos cada uno de estos aspectos, de momento demos la estructura básica de un MLG como en Wood [2017].

Definición 2.3.1 (Estructura básica de un MLG)**.**

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} E[Y_1] \\ \vdots \\ E[Y_n] \end{pmatrix} = E[Y]$$

$$g(\mu_i) = X_i\beta, \quad \forall i = 1, \ldots, n$$

Donde:

- X es la matriz modelo de dimensión $n \times p$ con $p \leqslant n$, que contiene a las variables predictoras, cada columna representa una variable predictora X_i.
- $\beta = (\beta_1, \ldots, \beta_p)^T$ es el vector de parámetros desconocidos como en el caso de los modelos lineales. A $\eta = X\beta$ se le conoce como predictor lineal.
- $g : \mathbb{R} \to \mathbb{R}$ es la función de enlace, que debe ser diferenciable y monótona. Representa la relación entre la media de las variables de respuesta y el predictor lineal.
- $Y = (Y_1, \ldots, Y_n)^T$ es un vector aleatorio, se suele suponer que las Y_i son variables aleatorias independientes y que siguen una distribución de tipo exponencial.

Observación 2.3.1. Desde esta formulación podemos ver fácilmente el por qué decimos que los MLG son una generalización de los modelos lineales, ya que basta tomar a la identidad como la función de enlace y suponer que la distribución considerada sea de tipo normal para encontrarnos ante la forma general de un modelo lineal como vimos en la sección anterior.

2.3.1. Familia de distribuciones exponenciales

Como hemos mencionado antes, la variable de respuesta de los modelos lineales generalizados deben seguir una distribución de tipo exponencial, en esta sección veremos qué significa eso y qué implicaciones tiene. Uno de los motivos más importantes por los que se supone que las variables de respuesta Y_i siguen distribuciones de esta familia se debe a que en los modelos lineales los cambios constantes en las variables predictoras implicaban cambios constantes en la variable de respuesta, pero ahora se quiere permitir que dichos cambios constantes de entrada puedan implicar también variaciones geométricas. Wikipedia [2023].

Definición 2.3.2 (Distribución de tipo exponencial)**.** Una distribución se dice que es de tipo exponencial si su función de densidad es de la forma:

$$f_\theta(y) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)}$$

Donde:

- a, b y c son funciones arbitrarias.
- ϕ es conocido como parámetro de escalado.
- θ es conocido como parámetro canónico de la distribución. Más adelante veremos que depende completamente de los parámetros del modelo β.

Ejemplo 2.3.1. La distribución normal es de tipo exponencial pues su función de densidad es:

$$f_\mu(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = e^{\frac{-y^2+2y\mu-mu^2}{2\sigma^2} - log(\sigma\sqrt{2\pi})} = e^{\frac{y\mu-\mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - log(\sigma\sqrt{2\pi})}$$

Y por tanto toma los siguientes parámetros de la familia exponencial:

- $\theta = \mu$
- $b(\theta) = \frac{\theta^2}{2} = \frac{\mu^2}{2}$
- $a(\phi) = \phi = \sigma^2$
- $c(\phi, y) = \frac{-y^2}{2\phi} - log(\sqrt{\phi 2\pi}) = \frac{-y^2}{2\sigma} - log(\sigma\sqrt{2\pi})$

Es posible dar una forma general para la esperanza y la varianza de las variables de tipo exponencial dependiendo de los parámetros de su función de densidad. Lo vemos en el siguiente resultado.

Proposición 2.3.1. Sea Y una variable de tipo exponencial, entonces verifica:

$$E[Y] = b'(\theta) \tag{2.4}$$

$$Var(Y) = b''(\theta)a(\phi) \tag{2.5}$$

Demostración. Partimos considerando la función de verosimilitud logáritmica para θ:

$$l(\theta) = log(f_\theta(y)) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

Y su derivada respecto de theta:

$$\frac{\partial l}{\partial \theta}(\theta) = \frac{y - b'(\theta)}{a(\phi)}$$

Ahora bien, si cambiamos la observación y por la variable Y, podemos evaluar la esperanza de esta derivada, la cual será 0 por propiedades de la función de verosimilitud logáritmica.

$$E[\frac{\partial l}{\partial \theta}(\theta)] = \frac{E[Y] - b'(\theta)}{a(\phi)} = 0$$

Y de aquí se obtiene directamente que $E[Y] = b'(\theta)$. Seguimos derivando para obtener que:

$$\frac{\partial^2 l}{\partial \theta^2}(\theta) = -\frac{b''(\theta)}{a(\phi)}$$

Y utilizando que: $E[\frac{\partial^2 l}{\partial \theta^2}] = -E[(\frac{\partial l}{\partial \theta})^2]$, nos queda que:

$$\frac{b''(\theta)}{a(\phi)} = \frac{E[(Y - b'(\theta))^2]}{a(\phi)^2} = \frac{E[(Y - E[Y])^2]}{a(\phi)^2} = \frac{Var(Y)}{a(\phi)^2}$$

De donde se obtiene que: $Var(Y) = b''(\theta)a(\phi)$. ■

La siguiente observación se basa en las indicaciones de Wood [2017].

Observación 2.3.2. Cuando ϕ es conocido el manejo de la función a no tiene dificultad, pero en muchos casos ϕ suele ser desconocido, así que para agilizar los resultados escribiremos $a(\phi) = \frac{\phi}{\omega}$, donde ω es una constante conocida. De hecho, todos los casos prácticos de interés se podrán expresar así y la mayoría con $\omega = 1$. De este modo nos queda: $Var(Y) = b''(\theta)\frac{\phi}{\omega}$. Por otro lado, en secciones posteriores necesitaremos trabajar con Var(Y) en función de $\mu = E[Y]$, para ello utilizaremos la relación 2.4 y definiremos una nueva función:

$$V(\mu) = \frac{b''(\theta)}{\omega} \tag{2.6}$$

pues de este modo se tiene que: $Var(Y) = V(\mu)\phi$.

Podemos recoger las carácterísticas de las principales distribuciones de tipo exponencial en la siguiente tabla:

	Binomial $Bi(n,p)$	Normal $N(\mu,\sigma^2)$	Poisson $Po(\lambda)$	Gamma $Ga(p,\lambda)$
$\theta(\mu)$	$log(\frac{\mu}{n-\mu})$	μ	$log(\mu)$	$-\frac{1}{\mu}$
ϕ	1	σ^2	1	$\frac{1}{p}$
$a(\phi)$	1	σ^2	1	$\frac{1}{p}$
$b(\theta)$	$nlog(1+e^\theta)$	$\frac{\theta^2}{2}$	e^θ	$-log(-\theta)$
$c(y,\phi)$	$log(\binom{n}{y})$	$-\frac{1}{2}(\frac{y^2}{\phi} + log(2\pi\phi))$	$-log(y!)$	$plog(py) - log(y\Gamma(p))$

2.3.2. Ajuste de los modelos lineales generalizados

La estimación de parámetros e inferencia con modelos aditivos generalizados se basa en la estimación de máxima verosimilitud, ya que, gracias a la hipótesis de que las Y_i pertenecezcan a la familia de distribuciones exponenciales, siempre se dispondrá de funciones de densidad. Partiremos considerando un MLG como el de la definición 2.3.1 y nuestro principal objetivo en esta sección será dar el estimador de máxima verosimilitud del vector de parámetros β como se hace en Wood [2017]. Veremos que será necesario recurrir a un algoritmo basado en mínimos cuadrados para hallar tal máximo, el método de mínimos cuadrados reponderados iterativamente.

Empezamos considerando y una observación de Y y notando que, como los Y_i son independientes entre sí, en tal caso la función de verosimilitud para β viene dada por:

$$L(\beta) = \prod_{i=1}^{n} f_{\theta_i}(y_i)$$

Y la función de verosimilitud logarítmica:

$$l(\beta) = \sum_{i=1}^{n} log(f_{\theta_i}(y_i)) = \sum_{i=1}^{n} \frac{y_i\theta_i - b_i(\theta_i)}{a_i(\phi)} + c_i(\phi, y_i)$$

La dependencia de β viene de que θ depende de μ y este a su vez depende del predictor lineal. Ahora bien, si sustituimos en esta expresión que $a_i(\phi) = \frac{\phi}{\omega_i}$, como mencionamos

en la sección anterior, nos queda:

$$l(\beta) = \sum_{i=1}^{n} \omega_i \frac{y_i\theta_i - b_i(\theta_i)}{\phi} + c_i(\phi, y_i)$$

Como queremos hallar el β que la hace máxima derivamos respecto β_i e igualamos a 0:

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \omega_i (y_i \frac{\partial \theta_i}{\partial \beta_j} - b_i'(\theta_i) \frac{\partial \theta_i}{\partial \beta_j})$$

Donde $\frac{\partial \theta_i}{\partial \beta_j}$ por la regla de la cadena es:

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{d\theta_i}{d\mu_i} \frac{\partial \mu_i}{\partial \beta_j} = \frac{d\theta_i}{d\mu_i} \frac{\partial \eta_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \eta_i} = \frac{X_{ij}}{g'(\mu_i) b''(\theta_i)}$$

Hemos utilizado que $\frac{\partial \mu_i}{\partial \eta_i} = g'(\mu_i)$, que $\frac{\partial \eta_i}{\partial \beta_j} = X_{ij}$ y que derivando el resultado 2.4 se tiene: $\frac{d\theta_i}{d\mu_i} = \frac{1}{b''(\theta)}$. Sustituyendo también el resultado de 2.6:

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{y_i - b_i'(\theta_i)}{g'(\mu_i) b_i''(\theta_i)/\omega_i} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{y_i - \mu_i}{g'(\mu_i) V(\mu_i)} X_{ij} \tag{2.7}$$

Observación 2.3.3. Ahora bien, el razonamiento general que seguiríamos sería el igualar estas derivadas a 0 y resolver el sistema resultante, sin embargo, se trata de un sistema con ecuaciones no lineales, por lo que para resolverlo utilizaremos métodos numéricos, en concreto utilizaremos el método de Newton que precisa del gradiente y del Hessiano de l. Por lo que debemos volver a derivar l.

$$\frac{\partial^2 l}{\partial \beta_j \beta_k} = -\frac{1}{\phi} \sum_{i=1}^{n} \frac{X_{ij} X_{ik} \alpha(\mu_i)}{g'(\mu_i)^2 V(\mu_i)}$$

Donde $\alpha(\mu_i) = 1 + (y_i - \mu_i)(\frac{V'(\mu_i)}{V(\mu_i)} + \frac{g''(\mu_i)}{g'(\mu_i)})$. Luego, si definimos la matriz $W = diag(\omega_i)$ para $\omega_i = \frac{\alpha(\mu_i)}{g'(\mu_i)^2 V(\mu_i)}$, el Hessiano de l es: $-\frac{1}{\phi} XWX$. De manera similar, si definimos $G = diag(\frac{g'(\mu)}{\alpha(\mu_i)})$, el gradiente de l puede escribirse como: $X^T W G \frac{(y-\mu)}{\phi}$.

De este modo, una actualización del método de Newton es de la forma:

$$\beta^{k+1} = \beta^k + (XWX)^{-1} X^T W G (y - \mu) = (XWX)^{-1} X^T W z$$

Donde $z_i = g'(\mu_i) \frac{y_i - \mu_i}{\alpha(\mu_i)} + \eta_i$

Observación 2.3.4. De lo anterior podemos notar que las actualizaciones de los β^k no son más que las estimaciones de β por mínimos cuadrados ponderados, es decir, resultan de minimizar:

$$\sum_{i=1}^{n} \omega_i (z_i - X_i \beta)^2 \tag{2.8}$$

Podemos entonces obtener el estimador numérico de los parámetros β de los MLG mediante el siguiente algoritmo.

Definición 2.3.3 (Algoritmo de mínimos cuadrados reponderados iterativamente)**.** 1) Inicialización: tomar $\hat{\mu_i} = y_i + \delta_i$ y $\hat{\eta_i} = g(\hat{\mu_i})$, donde δ_i suele ser 0 o una constante que asegure que $\hat{\eta_i}$ sea finito.
2) Calcular: $z_i = g'(\mu_i)\frac{y_i - \hat{\mu_i}}{\alpha(\mu_i)} + \hat{\eta_i}$ y $\omega_i = \frac{\alpha(\hat{\mu_i})}{g'(\hat{\mu_i})^2 V(\hat{\mu_i})}$
3) Encontrar $\hat{\beta}$ el parámetro que minimiza la función objetivo de mínimos cuadrados ponderados 2.8 y actualizar $\hat{\eta} = X\hat{\beta}$ y $\hat{\mu_i} = g^{-1}(\hat{\eta_i})$

Observación 2.3.5. Para saber con qué iteracion quedarnos debemos comprobar paso a paso si la variación entre el valor antiguo y el nuevo de $\hat{\beta}$ o las derivadas de la función de verosimilitud logáritmica están lo suficientemente cerca de 0.

2.3.3. Funciones de enlace canónicas

Definición 2.3.4 (Funciones de enlace canónicas)**.** Se dice que una función de enlace g_c es canónica para una distribución de la familia exponencial si verifica: $g_c(\mu_i) = \theta_i$, es decir, relaciona directamente el parámetro canónico con el predictor lineal.

Proposición 2.3.2 (Propiedades de las funciones de enlace canónicas)**.**

- Su uso en el modelo 2.3.1 resulta en: $\theta_i = X_i\beta$
- Hacen que $\alpha(\mu_i) = 1$.
- El Hessiano de la función de verosimilitud logarítmica coincida con su valor esperado.
- El sistema a resolver para obtener los estimadores de máxima verosimilitud de β, es decir, el formado por las derivadas parciales $\frac{\partial l}{\partial \beta_j}$ igualadas a 0 se reduce a: $X^t y - X^T\hat{\mu} = 0 \Rightarrow X^t y = X^T\hat{\mu}$ pues $\frac{\partial \theta_i}{\partial \beta_j} = X_{ij}$

2.3.4. Residuos

Al igual que para los modelos lineales, el estudio de los residuos ϵ_i de los MLG es un buen método para el control de los modelos, pero en este caso la estandarización de los residuos es necesaria y más complicada. Esto se debe a que si las suposiciones hechas sobre el modelo son correctas, entonces los residuos estandarizados deben tener aproximadamente la misma varizana y se deben comportar, tanto como sea posible, como los residuos de los modelos lineales. Para ello veremos dos tipos de estandarizaciones distintas.

Definición 2.3.5 (Residuos de Pearson)**.** Se dividen los residuos entre una cantidad proporcional a la desviación estándar dada por el modelo ajustado:

$$\hat{\epsilon}_i^p = \frac{y_i - \hat{\mu_i}}{\sqrt{V(\hat{\mu_i})}}$$

Deben tener media nula y varianza ϕ si el modelos es correcto.

Observación 2.3.6. Estos residuos no deberían mostrar ninguna tendencia en la media o la varianza al representarlos frente a los valores ajustados del modelo.

En la práctica los residuos de Pearson suelen ser asimétricos en torno al 0, lo que no concuerda mucho con que se parezcan a los residuos de los modelos lineales. Por tanto, se considera también la siguiente estandarización de los residuos, que surge al comparar la desviación del MLG con la suma de los residuos al cuadrado de los modelos lineales. Veámos primero a qué nos referimos con desviación:

Definición 2.3.6 (Desviación)**.** En el contexto de la definición 2.3.1, decimos que la desviación del modelo se define como:

$$D = 2(l(\hat{\beta}_{max}) - l(\hat{\beta})\phi = \sum_{i=1}^{n} 2\omega_i(y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i + b(\hat{\theta}_i)) \tag{2.9}$$

Donde:

- $l(\hat{\beta}_{max})$ representa el valor máximo de la función de verosimilitud logarítmica del modelo saturado (el que tiene un parámetro por dato). Este es el valor máximo que pueden tener todas las funciones de verosimilitud logarítmicas para los datos dados.
- $\tilde{\theta}_i$ es la estimación del parámetro canónico para el modelo saturado.
- $\hat{\theta}_i$ es la estimación del parámetro canónico del modelo que estamos estudiando.

Definición 2.3.7 (Residuos de desviación)**.** Si denotamos por d_i a la i-ésima componente de la desviación de un MLG, nos queda que $D = \sum_{i=1}^{n} d_i$ y se definen los residuos de desviación como:

$$\hat{\epsilon}_i^d = signo(y_i - \hat{\mu}_i)\sqrt{d_i}$$

Observación 2.3.7. Obviamente se tiene que: $D = \sum_{i=1}^{n}(\hat{\epsilon}_i^d)^2$. Además, notemos que: $D^* = \frac{D}{\phi} \backsim \chi_n^2$ y, aunque esto no se pueda trasladar directamente componente a componente, podemos intuir que:

$$\frac{d_i}{\phi} \backsim \chi_1^2 \Rightarrow \hat{\epsilon}_i^d \backsim N(0, \phi)$$

Es decir, intuitivamente, se comportarán como los residuos de los modelos lineales.

Capítulo 3

Modelos aditivos generalizados

3.1. Introducción

Como bien podemos intuir por su nombre, los modelos aditivos generalizados no son más que la fusión entre los modelos lineales generalizados y los modelos aditivos, los cuales se introducen con una sección en este capítulo. Podemos ver estos dos tipos de modelos como extensiones del modelo lineal. Por un lado, como vimos en el capítulo anterior, el MLG hace uso de una función de enlace entre el predictor lineal y el valor esperado de la variable dependiente para poder expresar relaciones más complejas y relaja la hipótesis distribucional permitiendo que tal variable siga distribuciones de la familia exponencial. Por otro lado, los modelos aditivos, además de también relajar esta hipótesis de distribución, introducen las funciones de suavizado en el modelo, estas proporcionan más flexibilidad a la hora de relacionar las variables explicativas con la de respuesta.

Luego, como ya hemos mencionado, y como se plantea en Hastie and Tibshirani [1990], el MAG reune estas dos propuestas de modo que generaliza el modelo aditivo de la misma forma que el MLG generalizaba el modelo lineal. Sin embargo, la flexibilidad que proporciona este modelo da lugar a dos nuevos problemas teóricos: cómo estimar las funciones de suavizado y cómo de "suaves" deben ser.

En este capítulo nos adentramos en los modelos no paramétricos, es decir, en aquellos que en vez de expresar la relación del valor esperado de la variable de respuesta con las variables predictoras mediante un predictor lineal, lo hacen mediante funciones f, como se vió en 2.1, pero ahora sin hacer ninguna suposición sobre ella. Esto conllevará en muchas ocasiones un mejor ajuste del modelo y traerá a la mesa una nueva cuestión conocida como sobreajuste que, aunque ya aparecía para los modelos paramétricos, ahora jugará un papel fundamental a la hora de querer predecir datos fuera de los observados. Este concepto refleja el hecho de que el modelo ajusta tan bien los datos proporcionados para la estimación de sus parámetros que es incapaz de mostrar la verdadera relación entre las variables que se estudian y, por tanto, da lugar a predicciones de nuevos datos que no serán las idóneas.

Tal y como se hace en Wood [2017], comenzaremos viendo cómo construir los modelos aditivos generalizados, es decir, qué bases de funciones podemos elegir para obtener las funciones de suavizado y qué parámetro de suavizado se debe seleccionar o cómo se puede estimar. Luego se introduce el modelo aditivo, en el que se utilizarán los resultados

vistos a lo largo del capítulo. Tras todo ello se propone la forma final del modelo aditivo generalizado.

Definición 3.1.1 (Estructura básica del modelo aditivo generalizado)**.**

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} E[Y_1] \\ \vdots \\ E[Y_n] \end{pmatrix} = E[Y]$$

$$g(\mu_i) = A_i\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \dots \quad , \forall i = 1, \dots, n \tag{3.1}$$

Donde:

- Y_i es la variable de respuesta y sigue una distribución de la familia exponencial de media μ_i y parámetro de escalado ϕ. A partir de ahora esto lo denotaremos por: $Y_i \backsim EF(\mu_i, \phi)$.
- A_i es la fila i-ésima de la matriz del modelo para aquellas componentes del modelo que son estrictamente paramétricas.
- θ es el correspondiente vector de parámetro, que antes denotábamos por β, para las variables predictoras mencionadas en el anterior punto.
- Las f_i son las funciones de suavizado para las covariables x_k. Suelen ser desconocidas y el principal objetivo es el de estimarlas, pero también pueden darse casos, la mayoría de modelos biológicos, en los que son conocidas y nos interesa estimar otros parámetros del modelo.

3.2. Suavizado univariante

Dicho esto, partiremos considerando modelos que, aunque no sean adecuados para un uso práctico general, nos permitirán estudiar el marco teórico de una forma más sencilla. Es decir, en esta sección consideraremos un modelo con una sola función de suavizado, f, y una sola covariable, x, de la forma:

$$y_i = f(x_i) + \epsilon_i \tag{3.2}$$

Donde y_i es la variable de respuesta y los ϵ_i son variables aleatorias independientes e identicamente distribuidas como $N(0, \sigma^2)$ que representan el error.

3.2.1. Bases de funciones

Nos proponemos en esta sección obtener una estimación de la función de suavizado a partir de una base de un espacio de funciones, en el que también se encontrará f (o una aproximación suya). Elegir una base equivale a tomar un conjunto de funciones $\{b_j(x)\}_{j=1}^k$ y, por tanto, podemos representar la función de suavizado como:

$$f(x) = \sum_{j=1}^{k} b_j(x)\beta_j \tag{3.3}$$

Para ciertos parámetros β_j a determinar.

Base polinómica
Si consideramos la base $\mathcal{B}$ del espacio de polinomios de grado k, es decir, $\mathcal{B} = \{1, x_i, x_i^2, \ldots, x_i^k\}$, la función de suavizado toma la forma:

$$f(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + \cdots + \beta_{k+1} x^k$$

Y, por tanto, el modelo 3.2 queda:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \cdots + \beta_{k+1} x_i^k + \epsilon_i$$

Observación 3.2.1 (Problema de la base polinómica). Notemos que por el teorema de Taylor, la base polinomial nos será útil cuando nuestro interés sea el de estudiar las propiedades de la función de suavizado en el entorno de un punto concreto, pero nos encontramos con problemas cuando queremos hacerlo en todo el dominio de f.

El principal problema se debe a que la interpolación de los datos puede resultar en una función muy oscilante o que no ajuste bien la información, dependiendo del valor de k, y que al modificar un coeficiente del modelo, el cambio impacta a los valores ajustados en todo el rango de la variable explicativa. Esto se puede solucionar de cierta manera con el siguiente tipo de base de funciones.

Base lineal por partes
Consideremos ahora una partición de nodos $\{x_j^* : j = 1, \ldots, k\}$ del rango de la variable predictora x tal que $x_j^* > x_{j+1}^*$ y la base de funciones $\mathcal{B} = \{b_j(x)\}_{j=1}^k$ donde:

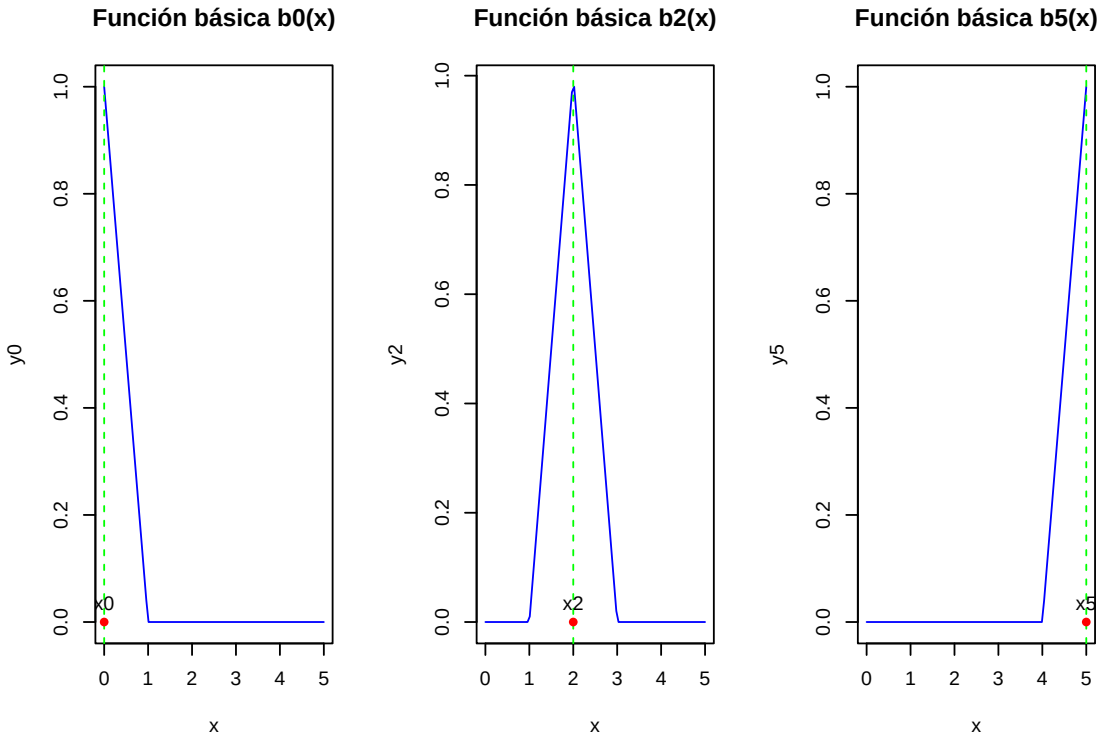
$$b_1(x) = \begin{cases} \frac{x_2^* - x}{x_2^* - x_1^*} & , si \;\; x < x_2^* \\ 0 \quad c.c. \end{cases}$$

$$b_j(x) = \begin{cases} \frac{x - x_{j-1}^*}{x_j^* - x_{j-1}^*} & , si \;\; x_{j-1}^* < x < x_j^* \\ \frac{x_{j+1}^* - x}{x_{j+1}^* - x_j^*} & , si \;\; x_j^* < x < x_{j+1}^* \\ 0 \quad c.c \end{cases}$$

$$b_k(x) = \begin{cases} \frac{x - x_{k-1}^*}{x_k^* - x_{k-1}^*} & , si \;\; x > x_{k-1}^* \\ 0 \quad c.c. \end{cases}$$

Es decir, la base de funciones $b_j(x)$ que son 0 en todo su dominio excepto entre los nodos a izquierda y derecha de x_j^*, donde crece y decrece de forma lineal hasta llegar a 1 en tal nodo. Este tipo de funciones se conocen como *tent functions*.

Ejemplo 3.2.1. Supongamos que el rango de x va de 0 a 5 y consideremos 6 nodos: $\{0, 1, 2, 3, 4, 5\}$, entonces podemos representar las funciones $b_0(x)$, $b_2(x)$ y $b_5(x)$ como:



De momento sólo planteamos estas formas de estimar las funciones de suavizado para tener una idea inicial y sencilla de cómo hacerlo pero más adelante dedicamos una sección a mejorar estas estimaciones mediante *splines*.

3.2.2. Control del suavizado

Nos interesará ahora controlar el grado de suavizado del GAM. Para ello tendremos en cuenta que el modelo aproxime de forma correcta los datos a la vez que la curvatura se mantiene controlada. Consideramos un nuevo parámetro λ, denominado parámetro de suavizado, el cuál tiene como principal función el compensar entre la fidelidad a los datos del modelos y el grado de suavizado del mismo.

Notemos primero que podemos representar la penalización a la curvatura de f como:

$$\int (f'')^2$$

Y en el caso de utilizar la base de funciones lineales por partes se puede aproximar[1] por:

$$\sum_{j=2}^{k-1} (f(x^*_{j-1}) - 2f(x^*_j) + f(x^*_{j+1}))^2$$

Es fácil observar que cuando f es una línea recta la penalización es 0 y cuando presenta muchas fluctuaciones en su curvatura este término es mayor.

[1]Se supone que se los nodos están espaciados de manera uniforme, pues en el caso de no que no lo estuvieran habría que añadir pesos a la suma.

Luego, en vez de ajustar el modelo por mínimos cuadrados, ahora se hará añadiendo la anterior penalización, es decir, minimizando:

$$||y - X\beta||^2 + \lambda \sum_{j=2}^{k-1} (f(x^*_{j-1}) - 2f(x^*_j) + f(x^*_{j+1}))^2 \tag{3.4}$$

Observación 3.2.2. Mientras mayor sea λ más importancia le estaremos dando a que la función f sea suave y menos a que a aproxime bien los datos. De hecho, cuando $\lambda \to \infty$ la función de suavizado f que minimiza la anterior expresión será una línea recta y cuando $\lambda = 0$ resultará en una estimación no penalizada.

3.2.3. Elección del parámetro de suavizado

Cómo hemos visto en la observación anterior: si el parámetro de suavizado es muy grande, el modelo será demasiado simple como para ajustarse bien a los datos y si es muy pequeño, la función de suavizado tendrá una curvatura muy alta. En cualquiera de los casos se tendrá que la estimación de f no se parecerá a la función real que ajusta los datos. Por ello, debemos dar un criterio para la elección de λ.

Un primer criterio planteado en Wood [2017] es el de elegir λ de forma que minimice la siguiente expresión para $x_1, \ldots, x_n$ unas observaciones dadas.

$$M = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_i - f_i)^2$$

Donde $\hat{f}_i = \hat{f}(x_i)$ es la evaluación de los puntos dados en la estimación de la función y $f_i = f(x_i)$ son sus evaluaciones en la función real.

Sin embargo, como la función f es desconocida, no es posible utilizar este criterio directamente. Daremos entonces una primera versión **método de validación cruzada**.

Definición 3.2.1 (Validación cruzada ordinaria)**.** Sea $\hat{f}_i^{[-i]}$ la estimación de la función de suavizado que ajustada por todos los datos $\{(x_j, y_j)\}_{j=1}^n$ menos el i-ésimo, se define la validación cruzada ordinaria como:

$$\nu_0 = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_i^{[-i]} - y_i)^2 \tag{3.5}$$

Se puede entender como que se ajusta el modelo sin utilizar la observación (x_i, y_i), se predice la variable de respuesta con este modelo en el punto x_i y luego se calcula la diferencia al cuadrado entre la estimación y el valor observado $\forall i = 1, \ldots, n$.

Observación 3.2.3. Podemos ver que tomar λ de modo que minimice ν_0 es una buena manera de abordar que minimice M. Para ello veamos que $E[\nu_0] \approx E[M] + \sigma^2$. Sustituyendo en 3.5 que $y_i = f_i + \epsilon_i$ nos queda que:

$$\nu_0 = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_i^{[-i]} - f_i + \epsilon_i)^2 = \frac{1}{n} \sum_{i=1}^{n} [(\hat{f}_i^{[-i]} - f_i)2 - 2(\hat{f}_i^{[-i]} - f_i)\epsilon_i + \epsilon_i^2]$$

Entonces, tomando valor esperado y teniendo en cuenta que $E[\epsilon_i] = 0$ y que ϵ_i y f_i son independientes:

$$E[\nu_0] = \frac{1}{n} E[\sum_{i=1}^{n} (\hat{f}_i^{[-i]} - f_i)^2] + \sigma^2 = E[M] + \sigma^2$$

Por lo tanto, cuando $n \to \infty$ se tienen las igualdades $E[\nu_0] = E[M] + \sigma^2$ y $\hat{f}^{[-i]} = \hat{f}$.

Si los modelos sólo fueran juzgados por su capacidad de ajustar los datos que les aportamos, entonces siempre se elegirían los modelos más complejos, pero el elegir el modelo que maximice la capacidad de predecir nuevos datos no tiene este problema.

Sin embargo, como se indica en Wood [2017], p.171, este método es costoso computacionalmente, ya que se deben realizar n ajustes de los datos, por ello se propone un nuevo método el cuál hace uso de la matriz de influencia A.

Definición 3.2.2 (Validación cruzada generalizada)**.** Dadas unas observaciones $\{(x_i, y_i)\}_{i=1}^n$ se elige λ tal que minimice:

$$\nu_g = n\frac{\sum_{i=1}^n (y_i - \hat{f}_i)^2}{(n - tr(A))^2}$$

3.3. Modelos aditivos

Como ya hemos mencionado previamente, el modelo aditivo es una extensión del modelo de regresión lineal. Su principal característica, la cual da lugar a su nombre, es que los efectos de las variables predictoras sobre la variable de respuesta son aditivos, es decir, una vez ajustado el modelo aditivo se pueden examinar tales efectors por separado. Veremos primero la forma general del modelo aditivo tal y como la introduce Hastie and Tibshirani [1990] y luego desarrollaremos la teoría alrededor de él para el caso de dos variables predictoras.

Definición 3.3.1 (Modelo aditivo)**.** Supongamos el contexto de las anteriores de las definiciones de modelos, el modelo aditivo se expresa como:

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon$$

Donde α es el término independiente, ϵ son errores aleatorios independientes de los X_j tales que $E[\epsilon] = 0$ y $Var(\epsilon) = \sigma^2$, y las f_j son funciones que conviene suponer univariadas y suaves pero no es necesario.

Observación 3.3.1. Además se debe tener que $E[f_j(X_j)] = 0 \quad \forall j = 1, \ldots, n$, pues de otro modo las funciones f_j añadirían términos independientes constantes adicionales.

Suele ser útil el pensar el modelo aditivo como un método que primero estima los parámetros adecuados en los que medir las variables y luego realiza el análisis lineal estándar sobre las variables transformadas. La principal motivación a priori tras este tipo de modelos es que, al representar por separado el efecto de cada variable predictora, mantienen la interpretabilidad del modelo lineal.

En lo que sigue, para poder ajustar más fácilmente el modelo como en Wood [2017], supondremos que se tienen sólo dos variables predictoras $X = (X_1, \ldots, X_p)$ y $V = (V_1, \ldots, V_p)$ y consideraremos el modelo aditivo:

$$y_i = \alpha + f_1(x_i) + f_2(v_i) + \epsilon_i \tag{3.6}$$

Observación 3.3.2. Los principales problemas del modelo aditivo son:

- La suposición de los efectos aditivos sobre 2.1 es bastante restrictiva.
- Existen problemas de identificabilidad pues las f_j son estimables con una precisión de una constante aditiva.

Sin embargo, si suponemos resueltos estos problemas, el modelo aditivo puede ser representado por splines de regresión penalizados, los cuales serán estimados mediante mínimos cuadrados penalizados, y el grado de suavizado, que se obtendrá por validación cruzada.

Modelo aditivo por regresión penalizada por partes

En lo que sigue, consideraremos la base del espacio de funciones lineales por partes vista en la sección anterior, es decir, expresamos las funciones f_1 y f_2 como:

$$f_1(x) = \sum_{j=1}^{k_1} b_j(x)\delta_j f_2(v) = \sum_{j=1}^{k_2} \beta_j(v)\gamma_j$$

Donde δ_j y γ_j son parámetros conocidos y las b_j y β_j son las funciones básicas de tipo carpa para los nodos x_j^* y v_j^* respectivamente, los cuales están espaciados uniformemente en el rango de x y v.

Definimos ahora los vectores n-dimesionales $\vec{f_1} = (f_1(x_1), \ldots, f_1(x_n))^T$ y $\vec{f_2} = (f_2(v_1), \ldots, f_2(v_n))^T$ como:

$$\vec{f_1} = X_1\delta = \begin{pmatrix} b_1(x_1) & \ldots & b_{k1}(x_1) \\ \vdots & \ldots & \vdots \\ b_1(x_n) & \ldots & b_{k1}(x_n) \end{pmatrix} \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_{k1} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{k1} b_j(x_1)\delta_j \\ \vdots \\ \sum_{j=1}^{k1} b_j(x_n)\delta_j \end{pmatrix} = \begin{pmatrix} f_1(x_1) \\ \vdots \\ f_1(x_n) \end{pmatrix}$$

$$\vec{f_2} = X_2\gamma = \begin{pmatrix} \beta_1(v_1) & \ldots & \beta_{k2}(v_1) \\ \vdots & \ldots & \vdots \\ \beta_1(v_n) & \ldots & \beta_{k2}(v_n) \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_{k1} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{k2} \beta_j(v_1)\gamma_j \\ \vdots \\ \sum_{j=1}^{k2} \beta_j(v_n)\gamma_j \end{pmatrix} = \begin{pmatrix} f_2(v_1) \\ \vdots \\ f_2(v_n) \end{pmatrix}$$

Proposición 3.3.1. En el caso de considerar la base lineal por partes, los coeficientes β_j que definen a una función f coinciden con los valores de la función en los nodos, es decir, $\beta_j = f(x_j^*)$.

Gracias a esto, se tiene que el problema de ajuste de la regresión penalizada se reduce a minimizar la siguiente expresión respecto de *beta*:

$$||y - X\beta||^2 + \lambda\beta^T S\beta$$

Donde $S = D^T D$ con $D = \begin{pmatrix} 1 & -2 & 1 & 0 & \ldots & \ldots \\ 0 & 1 & -2 & 1 & 0 & \ldots \\ 0 & 0 & 1 & -2 & 1 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ddots \end{pmatrix}$.

Por lo tanto, la penalización asociada a las funciones f_1 y f_2 vienen dadas por:

$$\begin{cases} \delta^T D_1^T D_1 \delta = \delta S_1 \delta \\ \gamma^T D_2^T D_2 \gamma = \gamma S_2 \gamma \end{cases}$$

Además, para tratar el problema de identificabilidad utilizaremos la siguiente restricción lineal:

$$\begin{cases} \sum_{i=1}^{n} f_1(x_i) = 0 \Leftrightarrow \bar{1}^T \vec{f_1} = 0 \Leftrightarrow \bar{1}^T X\delta = 0 \ \ \forall\delta \Leftrightarrow \bar{1}^T X = 0 \\ \sum_{i=1}^{n} f_2(v_i) = 0 \Leftrightarrow \bar{1}^T \vec{f_2} = 0 \Leftrightarrow \bar{1}^T V\gamma = 0 \ \ \forall\gamma \Leftrightarrow \bar{1}^T V = 0 \end{cases}$$

Donde $\bar{1}$ es un vector n-dimensional con todas las componentes iguales a 1. Ahora bien, para que se pueda cumplir esta condición debemos retirar de cada columna de las marices X y V la media de tales columnas, es decir, definiremos las nuevas matrices centradas por columnas y las respectivas transformaciones de f_1 y f_2:

$$\begin{cases} \tilde{X} = X - \bar{1}\bar{1}^T \frac{X}{n} \ \ , \ \ \tilde{f}_1 = \tilde{X}\delta \\ \tilde{V} = V - \bar{1}\bar{1}^T \frac{V}{n} \ \ , \ \ \tilde{f}_2 = \tilde{V}\gamma \end{cases}$$

Observación 3.3.3. Esta nueva restricción y la transformación de las funciones no afecta a las restricciones impuestas con anterioridad, de hecho solo implica un cambio constante en las funciones:

$$\tilde{f}_1 = \tilde{X}\delta = X\delta - \bar{1}\bar{1}X\frac{\delta}{n} = X\delta - \bar{1}c = f_1 - c$$

Para la constante $c = \bar{1}X\frac{\delta}{n}$.Se hace de forma análoga para f_2.

Finalmente, notemos que el proceso de centrado por columnas reduce el rango a $k_1 - 1$, así que sólo se podrán estimar $k_1 - 1$ de los k_1 elementos de δ de forma única. Para solucionar este problema se retira una columna de $\tilde{X}$ y de D_1 y la correspondiente componente de δ se hace 0.

Gracias a este razonamiento, el modelo aditivo puede ser expresado como $Y = Z\beta + \epsilon$, donde $Z = (\bar{1}, X, V)$ y $\beta = (\alpha, \delta, \gamma)^T$. De este modo, la penalización que añadimos al criterio de mínimos cuadrados es:

$$\beta^T S_1 \beta = (\alpha, \delta^T, \gamma^T) \begin{pmatrix} 0 & 0 & 0 \\ 0 & S_1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \delta \\ \gamma \end{pmatrix} = \delta^T S_1 \delta$$

Ajuste del modelo aditivo por mínimos cuadrados penalizados Gracias a la expresión de la penalización que acabamos de obtener para el modelo aditivo 3.6 se tiene que la estimación de los coeficientes $\hat{\beta}$ del modelo se obtienen minimizando la función objetivo de mínimos cuadrados penalizados:

$$||y - X\beta||^2 + \lambda_1 \beta^T S_1 \beta + \lambda_2 \beta^T S_2 \beta \tag{3.7}$$

Donde λ_1 es el parámetro de suavizado que controla a f_1 y λ_2 el que controla a f_2. Entonces, los estimadores de β son:

$$\hat{\beta} = (X^T X + \lambda_1 S_1 \lambda_2 S_2)^{-1} X^T Y$$

3.4. Smoothers

En esta sección razonaremos de forma similar a la sección 3.2.1, es decir, queremos definir una base de funciones $\mathcal{B} = \{b_i(x)\}_{i=1}^{k}$ de forma que las funciones de suavizado

se puedan expresar como una combinación lineal de estas funciones básicas. La base lineal por partes, vista en la sección antes mencionada, ofrece una forma razonable de representar las funciones de suavizado para los modelos aditivos, pero nos proponemos ahora dar mejores bases de funciones que tengan el mismo objetivo. Para ello, utilizaremos bases de splines ya que, para un tamaño de base fijado, reducen significativamente el error de aproximación de las funciones de suavizado.

3.4.1. Splines cúbicos

Será de interés el dedicarle una sección al marco teórico tras los splines, pues están muy relacionados con la mayoría de suavizadores. Como en Wood [2017], no abordaremos el tema de forma general, sino que se pueden recoger las ideas principales mediante las propiedades de los splines cúbicos. Veremos esto primero en el contexto de la interpolación y luego en el del suavizado.

Definición 3.4.1 (Spline cúbico)**.** Dada una colección de puntos $\mathcal{C} = \{(x_i, y_i)/i = 1, \ldots, n\}$ tales que $x_i \leq x_{i+1}$, decimos que $s : [x_1, x_n] \to \mathbb{R}$ es un spline cúbico si verifica:

- $s|_{[x_j, x_{j+1}]} = s_j$, donde s_j es un polinomio de grado 3 en $[x_j, x_{j+1}]$.
- $s(x_j) = y_j \quad \forall j = 1, \ldots, n.$
- s es continua hasta la segunda deriva en los nodos x_j, es decir, se cumple que:

$$s_{j+1}(x_{j+1}) = s_j(x_{j+1}), \quad s'_{j+1}(x_{j+1}) = s'_j(x_{j+1}), \quad s''_{j+1}(x_{j+1}) = s''_j(x_{j+1}) \ \forall j = 1, \ldots, n$$

Se dice que un spline cúbico es natural cuando: $s''(x_1) = 0 = s''(x_n)$

Splines cúbicos naturales como interpoladores

Partimos considerando una colección de puntos $\mathcal{C} = \{(x_i, y_i)/i = 1, \ldots, n\}$ tales que $x_i \leq x_{i+1}$ y el spline cúbico natural $s(x)$ que interpola los puntos de $\mathcal{C}$. Entonces, en Wood [2017] se propone la siguiente proposición como uno de los resultados más importantes en la teoría de splines:

Proposición 3.4.1. De entre todas las funciones f continuas en $[x_1, x_n]$, con primera derivada absolutamente continua y que interpolan los puntos de $\mathcal{C}$, $s(x)$ es la más suave de todas ellas en el sentido que minimiza:

$$J(f) = \int_{x_1}^{x_n} f''(x)^2 dx$$

Demostración. Sea $f(x)$ una función que verifique las condiciones del enunciado y sea distinta de $s(x)$, definimos $h(x) = f(x) - s(x)$ y busquemos una expresión de $J(f)$ en función de $J(s)$:

$$J(f) = \int_{x_1}^{x_n} f''(x)^2 dx = \int_{x_1}^{x_n} (s''(x)+h''(x))^2 dx = \int_{x_1}^{x_n} s''(x)^2 dx + 2\int_{x_1}^{x_n} s''(x)h''(x)dx + \int_{x_1}^{x_n} h''(x)^2 dx$$

Ahora bien, aplicando integración por partes en el término del medio de la última igualdad queda:

$$\int_{x_1}^{x_n} s''(x)h''(x)dx = s''(x_n)h'(x_n) - s''(x_1)h'(x_1) - \int_{x_1}^{x_n} s'''(x)h'(x)dx =$$

$$-\int_{x_1}^{x_n} s'''(x)h'(x)dx = -\sum_{i=1}^{n-1} s'''(x_i^+)\int_{x_i}^{x_{i+1}} h''(x)dx = -\sum_{i=1}^{n-1} s'''(x_i^+)(h(x_{i+1}) - h(x_i)) = 0$$

Hemos utilizado que $s''(x_1) = 0 = s''(x_n)$ y que como s es un spline cúbico, s''' es constante en cada intervalo (x_i, x_{i+1}), x_i^+ denota cualquier elemento de tal intervalo. Luego, nos queda que:

$$\int_{x_1}^{x_n} f''(x)^2dx = \int_{x_1}^{x_n} s''(x)^2 + \int_{x_1}^{x_n} h''(x)^2dx \geq \int_{x_1}^{x_n} s''(x)^2$$

De hecho, la igualdad sólo se tiene cuando $h''(x) = 0 \ \forall x \in (x_1, x_n)$. Sin embargo, como $h(x_1) = 0 = h(x_n)$ pues f y s tienen los mismos valores en los nodos, podemos observar que se da la igualdad si y sólo si $h(x) = 0 \ \forall x \in [x_1, x_n]$. Es decir, cualquier otra función de interpolación distinta de s tendrá mayor valor de la integral del cuadrado de su segunda derivada. ■

Esta proposición nos indica una razón de por qué el spline cúbico es el interpolador más suave para cualquier conjunto de datos. Sin embargo, esta no es la única propiedad interesante para tener en cuenta a los splines cúbicos a la hora de querer estimar las funciones de suavizado, como indica Wood [2017], en de Boor(1978, Capítulo 5), se enumeran una serie de resultados que indican que sea cual sea la verdadera función que representa los datos, un spline debe ser capaz de aproximarla de manera eficaz y además, si quisieramos construir un modelo a partir de funciones de suavizado de las covariables, el aproximar estas funciones por las aproximaciones más suaves puede ser una idea llamativa.

Splines cúbicos de suavizado

En los estudios estadísticos los datos con los que trabajamos suelen ser medidos con ruido, así que por lo general será más útil el suavizar los datos que el interpolarlos. Con tal propósito, nos será más conveniente el tratar a $s(x_i)$ como n parámetros de un spline cúbico, los cuáles pueden ser estimados minimizando:

$$\sum_{i=1}^{n}(y_i - s(x_i))^2 + \lambda \int s''(x)^2dx$$

Donde λ es el parámetro ajustable de suavizado. La función s resultante se conoce como Spline cúbico de suavizado y, de hecho, verifica:

Proposición 3.4.2. El spline cúbico de suavizado $s(x)$ minimiza:

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int f''(x)^2dx \tag{3.8}$$

para toda función f continua en $[x_1, x_n]$, con derivada absolutamente continua.

La prueba de este resultado se basa en la propiedad de los interpoladores 3.4.1.

Proposición 3.4.3. Se obtiene el mismo resultado de 3.4.2 si en vez de utilizar la suma de los cuadrados de los residuos en 3.8 utilizamos la log-verosimilitud.

Debido a estos resultados, los splines cúbicos parecen los suavizadores ideales pero tienen un problema computacional importante, ya que se deberán estimar tantos parámetros como datos queramos suavizar.

Observación 3.4.1 (Splines de regresión penalizada). Una buena forma de balancear las propiedades que ofrecen los splines con la eficiencia computacional será el utilizar mínimos cuadrados penalizados 3.7. En su forma más simple, esto conlleva formar una base de splines junto con sus respectivas penalizaciones para un conjunto de datos de cardinal menor al conjunto original y luego utilizarlos para modelar el conjunto completo.

Con este método nos surge la duda de cuántos elementos debe tener la base de splines. Generalmente esto no es posible saberlo sin conocer la función real que queremos estimar, pero es posible observar cómo debe escalar la dimensión de la base cuando se aumenta el número de datos. Se ve de manera detallada en @Wood (Capítulo 5.2, p.199).

3.4.2. Smoothers unidimensionales

Nos disponemos ahora a estudiar algunas bases penalizadas de suavizadores de las que se desarrollan en Wood [2017] en el caso de tener una sola variable predictora.

Splines cúbicos de regresión

Acabamos de ver varias propiedades interesantes de los splines cúbicos como suavizadores, veamos ahora como construirlo a partir de unos datos dados. Existen muchas bases equivalentes para representar splines cúbicos, nosotoros lo abordaremos como en Wood [2017], parametrizando el spline en términos de su valor en los nodos.

Partimos considerando un spline cúbico $f(x)$ y k nodos $\{x_1, \ldots, x_k\}$ con $x_i \leq x_{i+1}$.Denotaremos entonces $\beta_j = f(x_j)$ y $\delta_j = f''(x_j)$, pues de esta forma el spline f puede escribirse como:

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\delta_j + c_j^+(x)\delta_{j+1} \quad , \forall x \in [x_j, x_{j+1}]$$

Sea $h_j = x_{j+1} - x_j$, las funciones básicas a_j^-, a_j^+, c_j^- y c_j^+ vienen dadas por:

$$a_j^-(x) = \frac{x_{j+1} - x}{h_j} \qquad c_j^-(x) = \frac{1}{6}[\frac{(x_{j+1} - x)^3}{h_j} - h_j(x_{j+1} - x)]$$
$$a_j^+(x) = \frac{x - x_j}{h_j} \qquad c_j^+(x) = \frac{1}{6}[\frac{(x - x_j)^3}{h_j} - h_j(x - x_j)]$$

Ahora bien, para representar formalmente las condiciones que debe cumplir el spline: continuidad hasta la segunda derivada en los nodos x_j y que en los nodos extremos, x_1 y x_k, la segunda derivada sea nula, se utiliza:

$$B\delta^- = D\beta$$

donde $\delta^- = (\delta_2, \ldots, \delta_{k-1})^T$, $\delta_1 = 0 = \delta_k$ y B y D son definidas como:

$$\begin{cases} B_{i,i} = \frac{h_i + h_{i+1}}{3} \\ B_{i,i+1} = \frac{h_{i+1}}{6} \\ B_{i+1,i} = \frac{h_{i+1}}{6} \end{cases} \quad \forall i = 1, \ldots, k-3 \qquad \begin{cases} D_{i,i} = \frac{1}{h_i} \\ D_{i,i+1} = -\frac{1}{h_i} - \frac{1}{h_{i+1}} \\ D_{i,i+2} = \frac{1}{h_{i+1}} \end{cases} \quad \forall i = 1, \ldots, k-2$$

Luego, definiendo $F^- = B^{-1}D$ y $F = \begin{pmatrix} \bar{0} \\ F^- \\ \bar{0} \end{pmatrix}$, donde $\bar{0}$ es una fila de ceros, tenemos que $\delta = F\beta$ y de esta forma el spline se puede reescribir como:

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)F_j\beta + c_j^+(x)F_j\beta \quad , \forall x \in [x_j, x_{j+1}]$$

Y renombrando las funciones básicas de cierta forma nos queda finalmente que:

$$f(x) = \sum_{i=1}^{k} b_i(x)\beta_i$$

Observación 3.4.2. Gracias a esta nueva forma de expresar el spline mediante las funciones básicas podemos expresar la penalización de curvatura de la siguiente forma:

$$\int_{x_1}^{x_k} f''(x)^2 dx = \beta^T D^T B^{-1} D\beta = \beta^T S\beta$$

Donde $S = D^T B^{-1} D$ se conoce como la matriz de penalizado asociada a tal base.

Además de proporcionar parámetros directamente interpretables, la base de splines cúbicos no requiere de ningún re-escalado de las covariables, sólo de la elección de los nodos x_j.

P-splines

3.5. Modelos aditivos generalizados

Cómo ya hemos mencionado varias veces a lo largo del trabajo y como se indica en Wood [2017], en los MAG se quiere predecir algunas funcionasmonótonas que expresan la relación entre los predictores y el valor esperado de la variable de respuesta. Del mismo modo que para los MLG la variable de respuesta debe seguir una distribución de tipo exponencial, 2.3.2.

Mientras que el modelo aditivo se estimaba mediante mínimos cuadrados penalizados, el MAG se estirmará utilizando máxima verosimilitud penalizada, aunque en la práctica se utiliza un algoritmo de iteración de mínimos cuadrados penalizados (PIRLS). Partimos dando la definión general del modelo como lo hace Wood [2017]:

Definición 3.5.1 (Estructura general del modelo aditivo generalizado)**.** Sean:

- Y un vector aleatorio n-dimensional, cuyas componentes siguen una distribución de tipo exponencial $y_i \backsim EF(\mu_i, \phi)$ $\forall i = 1, \ldots, n$.

- X una matriz de orden $n \times p$, $p \leq n$, de columnas $x_j \ \forall j = 1, \ldots, p$ con constantes conocidas.
- $f_j : \mathbb{R} \to \mathbb{R} \ \forall j = 1, \ldots, p$ funciones desconocidas.
- A la matriz asociada al modelo paramétrico de orden $n \times p$.
- γ un vector de parámetros de dimensión n.
- $g : \mathbb{R} \to \mathbb{R}$ una función monótona y diferenciable.

Definimos el Modelo Aditivo Generalizado como:

$$g(\mu_i) = A_i\gamma + \sum_{j=1}^{p} f_j(x_{ij}) \tag{3.9}$$

Wood [2017] también propone una definición alternativa utilizando distintas formas de L_{ij}:

$$g(\mu_i) = A_i\gamma + \sum_{j=1}^{p} L_{ij} f_j(x_j)$$

La más básica de ella es: $L_{ij} f_j(x_j) = f_j(x_{ij})$ que da la forma habitual del MAG, pero existen otras variaciones de ellas y cada una tiene una utilidad específica.

Luego, para cada f_j se deben elegir una base de suavizado y una penalización, las cuales dan lugar a las matrices modelo $X^{[j]}$ y a la matriz de penalización $S^{[j]}$, de modo que $X^{[j]}_{ik} = b_{jk}(x_{ji})$ para b_{jk} la k-ésima función básica para f_j.

3.5.1. Ajuste del modelo

Seguiremos las directrices dadas en Wood [2017] para dar la estimación de los parámetros β cuando suponemos dado los parámetros de suavizado λ. Partimos notando que el vector de parámetros β contiene a γ y a

Capítulo 4

Título del Capítulo

4.1. Primera sección

Apéndice A

Apéndice: Título del Apéndice

A.1. Primera sección

Apéndice B

Apéndice: Título del Apéndice

B.1. Primera sección

Bibliografía

JJ Allaire, Yihui Xie, Christophe Dervieux, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *rmarkdown: Dynamic Documents for R*, 2023. URL https://CRAN.R-project.org/package=rmarkdown. R package version 2.21.

T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1990. ISBN 9780412343902. URL https://books.google.it/books?id=qa29r1Ze1coC.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York, 2014. ISBN 9781461471370. URL https://books.google.es/books?id=at1bmAEACAAJ.

Pedro L. Luque-Calvo. *Escribir un Trabajo Fin de Estudios con R Markdown*, 2017.

Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, 2023a. URL https://CRAN.R-project.org/package=ggplot2. R package version 3.4.4.

Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. *dplyr: A Grammar of Data Manipulation*, 2023b. URL https://CRAN.R-project.org/package=dplyr. R package version 1.1.1.

Wikipedia. Modelo lineal generalizado — wikipedia, la enciclopedia libre, 2023. URL https://es.wikipedia.org/w/index.php?title=Modelo_lineal_generalizado&oldid=151532785. [Internet; descargado 30-mayo-2023].

S.N. Wood. *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2017. ISBN 9781498728348. URL https://books.google.it/books?id=HL-PDwAAQBAJ.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2023. URL https://yihui.org/knitr/. R package version 1.42.