USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Trabajo Fin de Grado

Modelos de regresión aditivos

Marina Ramallo Blanco

2022/2023

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRADO DE MATEMÁTICAS

Trabajo Fin de Grado

Modelos de regresión aditivos

Marina Ramallo Blanco

Julio, 2023

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estatística e Investigación Operativa

Título: Modelos de regresión aditivos

Breve descripción del contenido

Los modelos de regresión nos permiten establecer la relación entre una variable respuesta y una o varias variables explicativas o también llamadas covariables. Habitualmente se considera que la relación entre la variable respuesta y las covariables se puede establecer en términos de un modelo paramétrico. Sin embargo, en muchas situaciones prácticas nos encontraremos con escenarios más complejos donde será necesario recurrir a modelos más flexibles como los modelos aditivos.

A modo de orientación, este trabajo podría organizarse en las siguientes secciones:

- Introducción a los modelos de regresión.
- Presentación de un modelo de regresión aditivo.
- Estimación del modelo a través del método de mínimos cuadrados penalizado.
- Selección del parámetro de suavizado.

Además, presentaremos diferentes modelos de regresión aditivos aplicados tanto a conjuntos de datos reales como a datos simulados. Para eso emplearemos el software estadístico libre R (https://www.r-project.org/).
Recomendaciones
Otras observaciones

Índice

Resumen

En muchas situaciones es de interés poder representar la relación de dependencia entre una variable respuesta y una o varias variables explicativas. Con este propósito introducimos los modelos de regresión. En una primera aproximación, lo más intuitivo es plantear un modelo de regresión paramétrico, es decir, que la forma del modelo sea totalmente conocida salvo por un cierto vector de parámetros, como es el caso de los modelos de regresión lineales. Sin embargo, en la práctica estos modelos no siempre ajustan bien la relación entre las variables que queremos representar, por lo que debemos recurrir a otro tipo de relaciones que nos den una mayor flexibilidad. En este contexto proponemos los modelos de regresión aditivos.

Los modelos de regresión aditivos son modelos no paramétricos, por lo que son muy flexibles, pero a la vez su formulación nos permite interpretar el efecto que tiene cada una de las variables explicativas sobre la variable respuesta. A lo largo de este manuscrito se presentarán métodos de estimación de los modelos aditivos (usando ideas de mínimos cuadrados) prestando especial interés a la elección de los parámetros de suavizado, como en cualquier modelo no paramétrico. Finalmente, se ilustrará la utilidad de los modelos aditivos empleando una base de datos reales.

Abstract

In many situations it is interesting to be able to represent the dependence relationship between a response variable and one or several explanatory variables. With this purpose we introduce regression models. In a first approximation, the most intuitive approach is to formulate a parametric regression model, that is, the form of the model would be completely known except for a certain parameter vector, as in the case of linear regression models. However, in practice these models not always adjust well the relationship between the variables we want to represent, so we must turn to other types of relationships which give us more flexibility. In this context we propose additive regression models.

Additive regression models are non-parametric models, so they are very flexible, but at the same time its formulation allows us to interpret the effect that each one of the explanatory variables has on the response variable. Throughout this manuscript estimation methods for additive models will be presented (using least squares ideas) paying special interest to the choice of the smoothing parameters, as in any non-parametric model. Finally, the utility of additive models will be ilustrated using a real database.

Capítulo 1

Introducción a los modelos de regresión

Los **modelos de regresión** son utilizados para explicar o modelar la relación entre una variable Y, llamada variable respuesta o variable dependiente, con respecto a un grupo de variables $X_1, \ldots, X_p$, llamadas variables explicativas o variables independientes.

Con la construcción de un modelo de regresión, podemos plantearnos dos objetivos fundamentales:

- Descripción de la relación de dependencia existente entre la variable respuesta y las variables explicativas.

- Predicción de futuras observaciones, es decir, predicción del valor que tomará la variable respuesta conocidos los valores de las variables explicativas.

Habitualmente, un modelo de regresión se formaliza como la **media condicionada** de la variable respuesta en función del valor que tomen las variables explicativas, es decir:

$$m(x_1, \ldots, x_n) = \mathbb{E}(Y|X_1 = x_1, \ldots, X_p = x_p)$$

para todos los posibles valores x_i de las variables X_i.

De este modo, en términos generales, un modelo de regresión adopta la forma:

$$Y = m(X_1, \ldots, X_p) + \varepsilon,$$

donde ε se define como el **error** del modelo, que es una variable aleatoria de media 0. Denominamos la función m como la **función de regresión**.

1.1. Regresión lineal múltiple

Dado que la función de regresión m es desconocida, resulta de interés tomar como hipótesis que m pertenezca a una familia paramétrica de funciones. Esto es lo que se conoce como **regresión paramétrica**. De este modo, la estimación del modelo se limitaría a estimar los valores de un número finito de parámetros en lugar de una función desconocida m, con lo que la dificultad del problema se reduciría notablemente.

Por ejemplo, podríamos suponer que m se trata de una función lineal, en cuyo caso hablaríamos de modelos de regresión lineales múltiples. Notemos que un modelo es lineal cuando se puede expresar como combinación lineal de los parámetros. Así, algunos ejemplos de modelos lineales serían:

$$\begin{aligned} Y =& \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, \\ Y =& \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 \log X_2 + \beta_4 X_1 X_2 + \varepsilon, \end{aligned}$$

mientras que el modelo siguiente sería no lineal:

$$Y = \beta_0 + e^{\beta_1 X_1} + \varepsilon.$$

1.1.1. El modelo de regresión lineal simple

Hablaremos de regresión lineal simple cuando tenemos una variable respuesta Y y una única variable explicativa X. El modelo de regresión lineal simple se basa en las siguientes hipótesis básicas:

- **Linealidad.** El modelo se escribe de la forma:

  $$Y = \beta_0 + \beta_1 X + \varepsilon,$$

  donde β_0 y β_1 son los parámetros del modelo, cuyos valores son desconocidos.

- **Homocedasticidad.** La varianza del error es la misma para cualquier valor de la variable explicativa:

  $$Var(\varepsilon | X = x) = \sigma^2 \text{ para todo } x.$$

- **Normalidad.** El error tiene distribución normal, esto es:

  $$\varepsilon \in N(0, \sigma^2).$$

- **Independencia.** Las variables aleatorias $\varepsilon_1, \dots \varepsilon_n$ son mutuamente independientes, donde n es el tamaño muestral.

1.1.2. El modelo de regresión lineal múltiple

Supongamos que tenemos una variable respuesta Y y un conjunto de variables explicativas $X_1, \ldots, X_p$. En este caso podríamos formular un modelo de regresión lineal múltiple como sigue:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

donde $\beta_0, \ldots, \beta_p$ serían los $p+1$ parámetros desconocidos del modelo y ε representaría el error del modelo.

Supongamos que tenemos una muestra de la forma:

$$\{(X_{1,1}, \ldots, X_{1,p}, Y_1), (X_{2,1}, \ldots, X_{2,p}, Y_2), \ldots, (X_{n,1}, \ldots, X_{n,p}, Y_n)\},$$

donde n es el número de observaciones. El modelo de regresión múltiple se podría expresar de la forma:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \varepsilon_i, \qquad i \in \{1, \ldots, n\}, \tag{1.1}$$

donde Y_i es el valor de la variable respuesta, $X_{i,1}, \ldots, X_{i,p}$ los valores de las covariables y ε_i el valor del error asociados al i-ésimo individuo de la muestra.

Al igual que en el modelo de regresión lineal simple, en el caso múltiple también supondremos que el error verifica las hipótesis de homocedasticidad, normalidad e independencia, es decir:

$$\varepsilon_1, \ldots, \varepsilon_n \in N(0, \sigma^2) \quad \text{y son independientes.}$$

El modelo de regresión lineal múltiple, tal y como fue formulado en la expresión (1.1), se puede expresar en notación matricial como sigue:

$$\mathbb{Y} = \mathbb{X}\beta + \varepsilon,$$

donde $\mathbb{Y}$ es el vector de valores observados de la variable respuesta, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ es el vector de parámetros y ε es el vector de errores. $\mathbb{X}$ es una matriz de dimensión $n \times (p+1)$ denominada **matriz de diseño**. En cada una de sus filas está representado un individuo, y en cada una de sus columnas los valores de una variable explicativa, excepto la primera columna, que consiste en una columna de unos que permite incorporar el término del intercepto β_0. Así, tenemos:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

1.2. Estimación del modelo lineal múltiple

Con la finalidad de estimar el modelo de regresión lineal múltiple, $\mathbb{Y} = \mathbb{X}\beta + \varepsilon$, tendremos que dar una estimación del vector de parámetros β y de la varianza del error σ^2.

1.2.1. Estimación del vector de parámetros

Supongamos que tomamos como estimador de β el vector $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)'$. Denotamos por $X_{i,1}, \ldots, X_{i,p}$ a los valores de las variables explicativas asociados al i-ésimo individuo de la muestra. Entonces, la predicción del valor de la variable respuesta proporcionada por el modelo será:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{i,1} + \cdots + \widehat{\beta}_p X_{i,p} = X_i \widehat{\beta}.$$

Sin embargo, el valor de la variable respuesta observado es Y_i. Definimos los **residuos** de un modelo de regresión como la diferencia entre las observaciones y las predicciones, es decir:

$$\widehat{\varepsilon}_i = Y_i - X_i\widehat{\beta} = Y_i - \widehat{Y}_i, \quad i \in \{1 \ldots, n\}.$$

Para la estimación de β plantearemos el **método de mínimos cuadrados**, que consiste en tomar como estimador de β aquel vector $\widehat{\beta}$ que dé lugar a los residuos más pequeños, es decir:

$$\widehat{\beta} = \arg\min_{\tilde{\beta}} \sum_{i=1}^{n} (Y_i - X_i\tilde{\beta})^2 = \arg\min_{\tilde{\beta}} (\mathbb{Y} - \mathbb{X}\tilde{\beta})'(\mathbb{Y} - \mathbb{X}\tilde{\beta}).$$

Derivando esta expresión respecto a $\tilde{\beta}$ e igualando a cero, obtenemos las **ecuaciones normales de regresión**:

$$\mathbb{X}'\mathbb{X}\widehat{\beta} = \mathbb{X}'\mathbb{Y},$$

cuya solución es el estimador de mínimos cuadrados $\widehat{\beta}$. Suponiendo que la matriz $\mathbb{X}'\mathbb{X}$ es invertible, este vendrá dado por:

$$\widehat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}.$$

Una vez hemos obtenido las estimaciones de los parámetros del modelo por mínimos cuadrados, podemos ver que los ajustes del modelo serían de la forma:

$$\widehat{\mathbb{Y}} = \mathbb{X}\widehat{\beta} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y} = H\mathbb{Y}, \tag{1.2}$$

donde la matriz $H = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$ se denomina **matriz *hat***. De este modo, los ajustes $\widehat{\mathbb{Y}}$ se obtienen aplicando la matriz H a las observaciones $\mathbb{Y}$.

Del mismo modo, los residuos serían de la forma:

$$\widehat{\varepsilon} = \mathbb{Y} - \widehat{\mathbb{Y}} = (I_n - H)\mathbb{Y} = M\mathbb{Y},$$

donde la matriz $M = (I_n - H)$ se conoce como **matriz generadora de residuos**.

Veamos una interpretación geométrica del método de mínimos cuadrados. Nuestro objetivo a la hora de estimar el vector de parámetros β del modelo es escoger un estimador $\widehat{\beta}$ tal que los ajustes del modelo $\widehat{\mathbb{Y}} = \mathbb{X}\widehat{\beta}$ estén lo más cerca posible de las observaciones $\mathbb{Y}$, o lo que es lo mismo, tal que se minimicen los residuos $\widehat{\varepsilon}$. Así, deberemos tomar $\widehat{\mathbb{Y}}$ como aquella combinación lineal de las columnas de $\mathbb{X}$ que minimice la distancia a $\mathbb{Y}$.

El vector $\widehat{\beta}$ que buscamos es, por lo tanto, aquel tal que $\widehat{\mathbb{Y}} \in \mathbb{R}^n$ sea la proyección ortogonal de $Y \in \mathbb{R}^n$ sobre el espacio $(p+1)$-dimensional generado por las columnas de $\mathbb{X}$, como podemos ver en la Figura 1.1. La estimación del vector de parámetros que acabamos de describir gráficamente se corresponde con la aplicación del método de mínimos cuadrados.

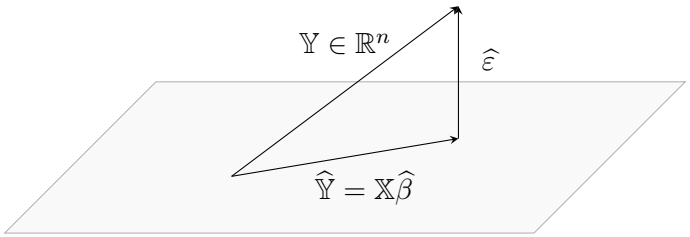


Figura 1.1: Interpretación geométrica del método de mínimos cuadrados.

Tal y como vimos en la expresión (1.2), el vector de ajustes se obtiene aplicando la matriz *hat* H al vector $\mathbb{Y}$. Como consecuencia, H se puede interpretar como la matriz asociada a la proyección sobre el espacio generado por las columnas de $\mathbb{X}$.

1.2.2. Estimación de la varianza

Dado que los errores no se observan, para la estimación de su varianza σ^2 emplearemos los residuos. Definimos la suma residual de cuadrados como:

$$RSS = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \widehat{\varepsilon}'\widehat{\varepsilon}.$$

Así, tomamos como estimador de la varianza del error:

$$\widehat{\sigma}^2 = \frac{RSS}{n - (p+1)},$$

donde el número de parámetros del modelo es $p + 1$.

1.2.3. Propiedades de los estimadores

Bajo las hipótesis básicas de un modelo lineal múltiple, es sencillo probar que el estimador $\widehat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}$ es insesgado y tiene varianza $(\mathbb{X}'\mathbb{X})^{-1}\sigma^2$. Además, por el teorema de Gauss-Markov sabemos que este estimador es el que tiene menor varianza entre los estimadores lineales insesgados de β.

También podemos ver que $\widehat{\sigma}^2 = \frac{RSS}{n-(p+1)}$ se trata de un estimador insesgado de σ^2. Bastaría con comprobar que la esperanza de RSS vale $\sigma^2(n-(p+1))$.

Bajo las hipótesis del modelo lineal múltiple, como consecuencia del Teorema de Fisher, las distribuciones en el muestreo que siguen los estimadores considerados anteriormente son las siguientes:

- $\widehat{\beta} \in N_{p+1}\left(\beta, \sigma^2(\mathbb{X}'\mathbb{X})^{-1}\right)$,
- $\dfrac{RSS}{\sigma^2} = \dfrac{(n-(p+1))\,\widehat{\sigma}^2}{\sigma^2} \in \chi^2_{n-(p+1)}$,

y además $\widehat{\beta}$ y $\widehat{\sigma}^2$ son independientes. Para profundizar en el estudio del modelo de regresión lineal múltiple, véase Faraway (2005).

1.3. Introducción a los modelos de regresión aditivos

Consideremos la variable respuesta Y y las variables explicativas $X_1, \ldots, X_p$. Para representar la relación entre ellas podríamos pensar en plantear un modelo de regresión lineal múltiple como hemos estudiado a lo largo de este primer capítulo. Sin embargo, hay multitud de situaciones prácticas en las cuales la hipótesis de linealidad no se cumple, con lo que el planteamiento del modelo lineal múltiple no sería correcto.

Alternativamente, podríamos considerar un modelo lineal tomando transformaciones de las variables explicativas o incluyendo interacciones entre ellas. Sin embargo, puede resultar muy complejo dar con el modelo adecuado debido al elevado número de posibilidades o incluso la transformación de variables podría no resultar suficiente.

En estos casos podemos pensar en plantear un **modelo no paramétrico** de la forma:

$$Y = m(X_1, \ldots, X_p) + \varepsilon, \tag{1.3}$$

donde la función m es totalmente desconocida.

En el Capítulo 2 estudiaremos este tipo de modelos. Como veremos, los modelos no paramétricos tienen ajustes bastante complejos, y para un número elevado de variables explicativas no

resulta sencilla su aplicación en la práctica.

Con el propósito de simplificar el modelo no paramétrico dado en (1.3), introducimos los **modelos de regresión aditivos**, que serán de la forma:

$$Y = \beta_0 + m_1(X_1) + m_2(X_2) + \cdots + m_p(X_p) + \varepsilon = \beta_0 + \sum_{j=1}^{p} m_j(X_j) + \varepsilon. \tag{1.4}$$

Los modelos aditivos nos dan mucha más flexibilidad que los lineales, pero mantieniendo la interpretabilidad de los efectos de las variables explicativas, ya que cada función m_i puede ser representada para darnos una idea de la relación marginal entre la respectiva covariable X_i y la variable respuesta. Además, la estimación de las funciones univariantes m_i asociadas al modelo (1.4) es notablemente más sencilla que la estimación de la función m asociada al modelo (1.3).

Ejemplo 1.1. Vamos a considerar un modelo de regresión aditivo de la siguiente forma:

$$Y = m(X_1, X_2, X_3, X_4) + \varepsilon = m_1(X_1) + m_2(X_2) + m_3(X_3) + m_4(X_4) + \varepsilon, \tag{1.5}$$

donde $\varepsilon \in N(0,1)$ y los efectos de cada variable explicativa sobre la variable respuesta vienen dados por:

$$\begin{aligned} m_1(X_1) =& 5X_1, \\ m_2(X_2) =& 1 - 48X_2 + 218X_2^2 - 315X_2^3 + 145X_2^4, \\ m_3(X_3) =& \sin(5\pi X_3), \\ m_4(X_4) =& 10(X_4^4 + X_4^2 - X_4), \end{aligned}$$

y podemos ver su representación en la Figura 1.2, suponiendo que las variables X_i siguen una distribución uniforme en el intervalo $[0,1]$.

Vamos a simular datos a partir del modelo (1.5). Con este propósito, para $i \in \{1, \ldots, 200\}$, generamos $X_{i,1}$, $X_{i,2}$, $X_{i,3}$ y $X_{i,4}$ de forma aleatoria de una distribución uniforme en el intervalo $[0,1]$ y los errores ε_i de una distribución normal de media 0 y varianza 1. Entonces, tendríamos que:

$$Y_i = m_1(X_{i,1}) + m_2(X_{i,2}) + m_3(X_{i,3}) + m_4(X_{i,4}) + \varepsilon_i,$$

con $i \in \{1, \ldots, 200\}$. De esta forma, obtenemos el conjunto de datos $\{(X_1, Y_1), \ldots, (X_{200}, Y_{200})\}$ simulados a partir del modelo (1.5).

Por lo general, si consideramos unos datos reales la función de regresión será desconocida. Es por esto que emplearemos datos simulados, que nos permitan ver en qué medida se aproxima la estimación del modelo a la función de regresión teórica.

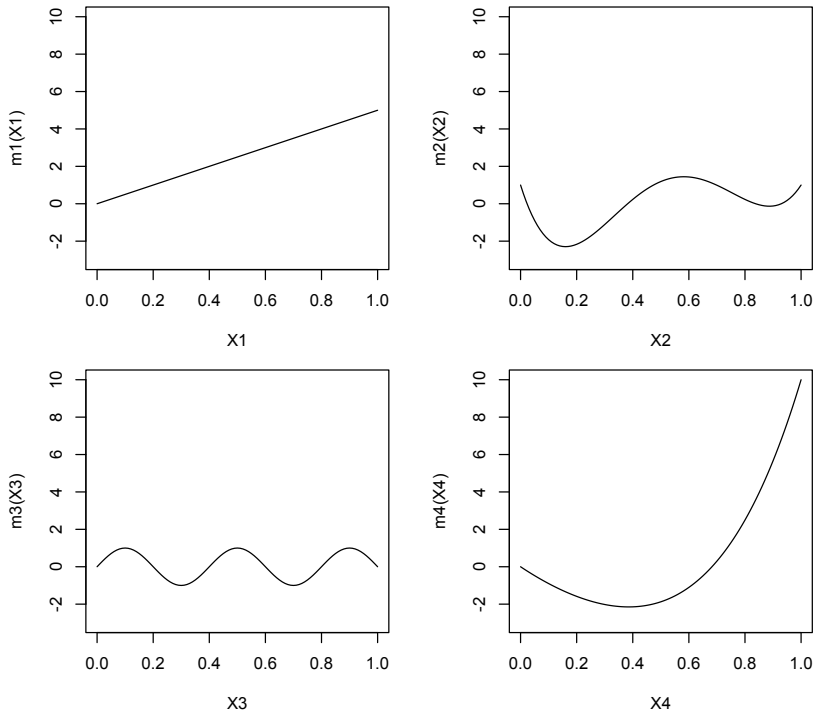


Figura 1.2: Efecto de cada una de las variables explicativas asociados al modelo de regresión aditivo dado en (1.5).

Considerando estos datos, procedemos a ajustar un modelo lineal múltiple como hemos visto a lo largo de este primer capítulo. Entonces tendríamos el modelo ajustado:

$$\widehat{Y} = -5.415 + 4.960X_1 + 2.153X_2 + 1.845X_3 + 8.289X_4.$$

Una vez ajustado el modelo nos planteamos la validación del mismo. Entonces, si observamos el primer gráfico de la Figura 1.3, que representa los residuos frente a los ajustes del modelo, podemos ver que tanto el valor de los residuos como su dispersión son mayores en los extremos y menores en el centro, contradiciendo así las hipótesis de linealidad y de igualdad de varianzas del modelo lineal múltiple. En la Figura 1.4 representamos los valores de la variable respuesta frente a los ajustes del modelo lineal múltiple. Como podemos observar, los datos no se distribuyen uniformemente en torno a la diagonal, que está representada por una línea discontinua. Por el contrario, los datos se concentran por encima de esta para los valores extremos y por debajo para los valores centrales. Esto indica un incumplimiento de las hipótesis del modelo.

En efecto, sabemos que la verdadera función de regresión viene dada por (1.5). De este modo, ya conocíamos de antemano que el modelo de regresión lineal múltiple no sería correcto para representar estos datos.

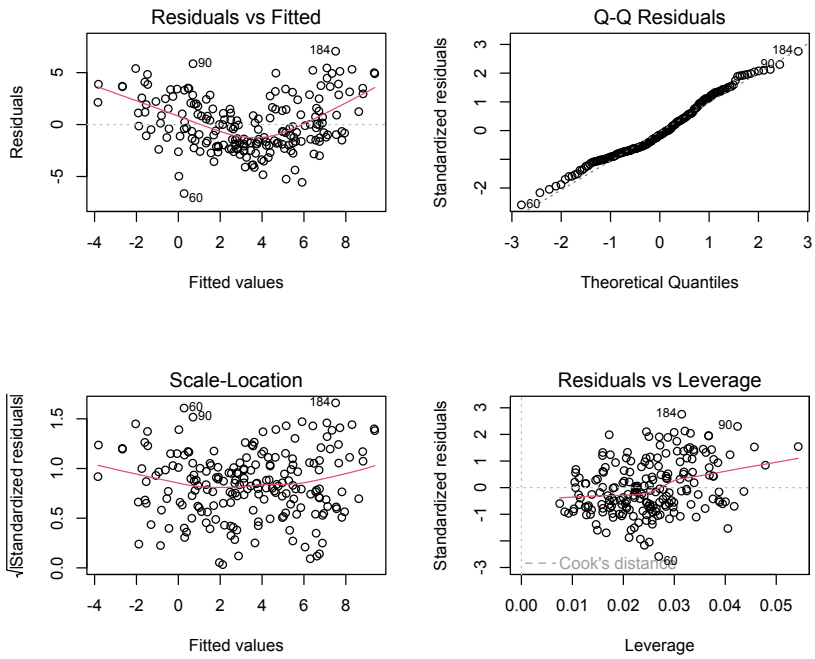


Figura 1.3: Gráficos para la validación del modelo de regresión lineal múltiple, obtenidos con la función `plot` del software estadístico R.

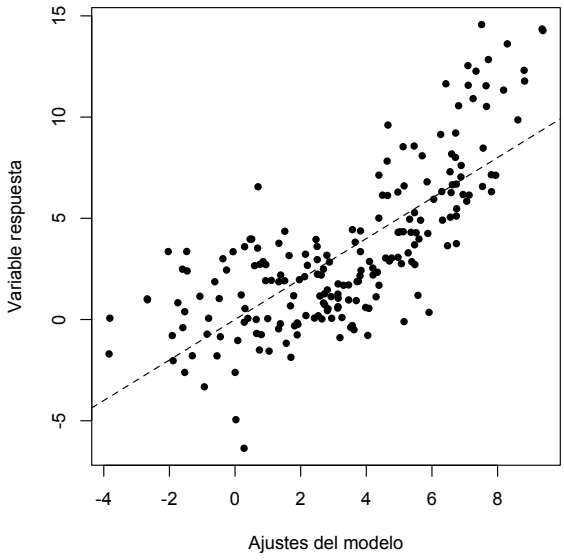


Figura 1.4: Representación de los valores reales frente a los valores ajustados con el modelo lineal múltiple.

Capítulo 2

Regresión no paramétrica

A lo largo de este capítulo estudiaremos los modelos de regresión no paramétrica en el contexto de la regresión simple, es decir, consideraremos una única variable explicativa X. Nos basaremos principalmente en Fan y Gijbels (1996).

Dada una muestra aleatoria simple de tamaño n, que denotaremos por $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, planteamos un modelo de la forma:

$$Y_i = m(X_i) + \varepsilon_i, \quad i \in \{1, \ldots, n\},$$

donde m es una función de regresión desconocida y ε_i son los errores del modelo.

Como vimos al inicio de la Sección 1.1, un modelo paramétrico se basa en suponer que la función m tiene una forma determinada, dependiendo únicamente de un número finito de parámeros desconocidos. En el caso no paramétrico, por el contrario, sólo tomaremos como hipótesis que m verifique ciertas condiciones de regularidad, sin asumir una forma específica para la función de regresión. En este sentido, los modelos no paramétricos nos dan mucha más flexibilidad, aunque también perdemos la interpretabilidad relativa a los parámetros asociados a un modelo paramétrico.

Dado un modelo de regresión no paramétrico, nuestro objetivo será estimar la función de regresión m. Para ello, existen diversas aproximaciones como la estimación constante local y la estimación lineal local, que veremos a continuación.

2.1. Regresión constante local

Una forma intuitiva de dar una estimación de $m(x)$, es decir, de la media condicional de la variable respuesta Y para $X = x$, es utilizando una media local ponderada en el punto x. Así,

tenemos el denominado **estimador de Nadaraya-Watson**, que viene dado por:

$$\widehat{m}_{NW,h}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right)},$$

donde la función K es una **función núcleo** o *kernel*, es decir, una función de densidad simétrica centrada en el 0. El parámetro h se denomina **parámetro ventana** o parámetro de suavizado, cuya elección resultará crucial para poder obtener un buen estimador. Como podemos observar, la estimación constante local proporciona una estimación para la media condicional m que es localmente constante en cada x.

De esta forma, a la hora de implementar un estimador tipo núcleo tendremos que llevar a cabo la elección de la función núcleo y del parámetro ventana, aunque la elección de este último resultará de mayor relevancia.

2.2. Regresión lineal local

La idea principal de los modelos de regresión lineal local, como su propio nombre indica, consiste en ajustar un modelo lineal en un entorno de cada posible valor de la variable explicativa. De este modo, dado un determinado parámetro ventana h construiremos un modelo lineal simple en el intervalo $(x - h, x + h)$ de la forma:

$$Y_i = \beta_0(x) + \beta_1(x)X_i + \varepsilon_i, \quad X_i \in (x - h, x + h).$$

A la vista del modelo anterior, podemos obtener los estimadores de los parámetros del modelo mediante el método de mínimos cuadrados ponderado, que consiste en tomar como estimadores de β_0 y β_1 aquellos que resuelvan el problema de minimización siguiente:

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg \min_{\tilde{\beta}_o, \tilde{\beta}_1} \sum_{i=1}^{n} \left(Y_i - \tilde{\beta}_0(x) - \tilde{\beta}_1(x)(x - X_i)\right)^2 \; K_h(x - X_i),$$

donde $K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$. Como podemos ver, se trata de un método de mínimos cuadrados en el que la introducción de la función K hace que los puntos cercanos a x tengan más influencia en el ajuste. Esto nos sugiere que se tienen las siguientes estimaciones:

$$\widehat{m}_h(x) = \widehat{\beta}_0$$
$$\widehat{m'_h}(x) = \widehat{\beta}_1$$

donde m' denota la derivada de la función de regresión respecto de x.

Dada K una determinada función núcleo y h un determinado parámetro ventana, en la Tabla 2.1, que podemos encontrar en Fan (1992), se resume el comportamiento asintótico del estimador de Nadaraya-Watson y del lineal local para una determinada función núcleo K y un parámetro de suavizado h.

Estimador	Sesgo	Varianza
Nadaraya-Watson	$\frac{1}{2}h^2\left(m''(x)+\frac{2m'(x)f'(x)}{f(x)}\right)\int_{-\infty}^{\infty}u^2K(u)du$	$\frac{\sigma^2(x)}{f(x)nh}\int_{-\infty}^{\infty}K^2(u)du$
Lineal local	$\frac{1}{2}h^2m''(x)\int_{-\infty}^{\infty}u^2K(u)du$	$\frac{\sigma^2(x)}{f(x)nh}\int_{-\infty}^{\infty}K^2(u)du$

Tabla 2.1: Propiedades asintóticas de los estimadores constante y lineal local.

A la vista de la Tabla 2.1, ambos estimadores tienen la misma varianza asintótica. Sin embargo, el estimador de Nadaraya-Watson tiene un alto sesgo especialmente en las regiones donde la derivada de la función de regresión $m'(x)$ toma un valor alto. Así, incluso en el caso en el que la función de regresión sea lineal, este estimador podría tener un sesgo elevado. Tampoco se adapta a diseños no uniformes, pues el sesgo puede ser alto cuando lo sea $\frac{f'(x)}{f(x)}$, siendo f la función de densidad de la variable explicativa.

La discrepancia entre el orden de magnitud del sesgo en el interior y cerca de la frontera es lo que se conoce como **efecto frontera**. El estimador de Nadaraya-Watson se ve más afectado por el efecto frontera que el estimador lineal local, que se adapta automáticamente a los puntos frontera (véase Fan y Gijbels (1996)).

En consecuencia, en la práctica suele ser preferible el empleo de un estimador lineal local. Es por esto que a continuación vamos a estudiar la selección de la función núcleo y del parámetro de suavizado exclusivamente para el caso lineal local.

2.2.1. Selección de la función núcleo

La elección de la función núcleo K no tiene demasiada relevancia, en el sentido de que no tiene un efecto tan crucial como el del parámetro ventana h en la bondad del estimador resultante. Algunos ejemplos de núcleos usados habitualmente son:

- Uniforme: $K(u)=\mathbb{I}(|u|<1)/2$,
- Gaussiano: $K(u)=\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-u^2}{2}\right)$,
- Epanechnikov: $K(u)=\frac{3}{4}(1-u^2)\ \mathbb{I}\{|u|<1\}$,

donde $\mathbb{I}$ representa la función indicadora. Como puede verse en la Sección 2.7 de Wand y Jones (1995), el núcleo de Epanechnikov puede considerarse óptimo, por lo que es recomendable su uso. Además, su soporte compacto permite una rápida computación.

2.2.2. Selección del parámetro de suavizado

La elección del parámetro ventana es crucial, pues una pequeña variación de este parámetro puede suponer importantes cambios en el estimador $\widehat{m}_h$. Tal y como se recoge en la Tabla 2.1, según aumentamos el parámetro ventana, aumentará el sesgo del correspondiente estimador, pero su varianza disminuirá. De este modo, si tomamos un parámetro ventana demasiado pequeño, aunque el sesgo del estimador sería pequeño, su varianza sería demasiado alta, dando lugar a un fenómeno que se conoce como **infrasuavizado**. Así, el estimador se aproximaría a recorrer todos los puntos de la muestra. Por el contrario, si tomamos un parámetro ventana demasiado grande, la varianza del estimador sería pequeña pero el sesgo muy grande, produciéndose lo que denominamos **sobresuavizado**, por lo que la estimación sería demasiado suave, pudiendo pasar por alto características relevantes de la muestra.

Ejemplo 2.1. Con la misma filosofía que introducimos en el Ejemplo 1.1, consideremos el siguiente modelo de regresión univariante:

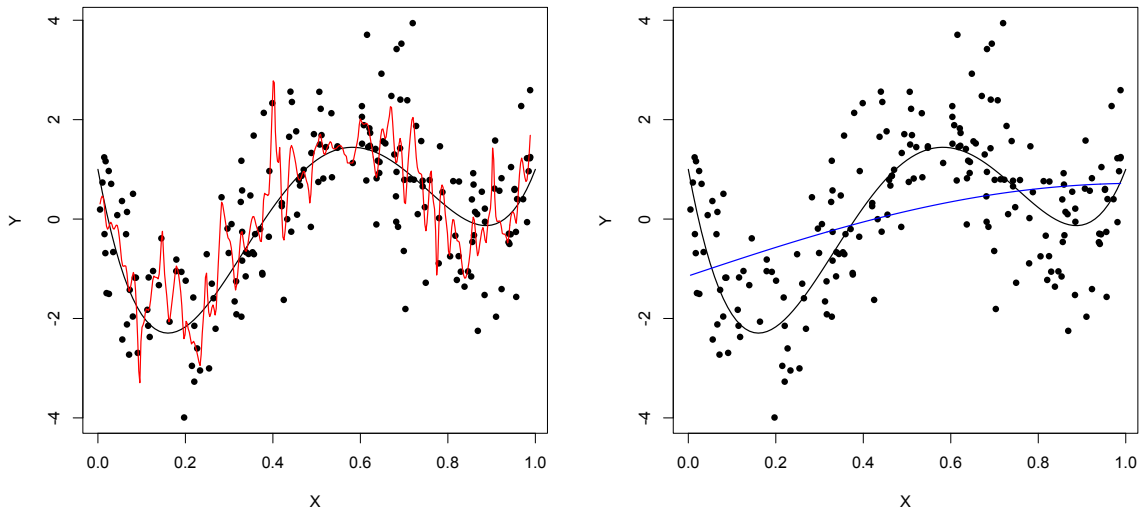
$$Y = m(X) + \varepsilon = 1 - 48X + 218X^2 - 315X^3 + 145X^4 + \varepsilon. \tag{2.1}$$

Ahora, del mismo modo que procedimos para el Ejemplo 1.1, obtenemos un conjunto de datos $\{(X_1, Y_1), \ldots, (X_{200}, Y_{200})\}$ simulados a partir de este modelo. Es decir, para $i \in \{1, \ldots, 200\}$ tomamos X_i de forma aleatoria de una distribución uniforme $U[0, 1]$ y los errores ε_i de una normal estándar. Entonces,

$$Y_i = 1 - 48X_i + 218X_i^2 - 315X_i^3 + 145X_i^4 + \varepsilon_i.$$

Ajustamos un modelo lineal local para estos datos tomando distintos parámetros de suavizado, como podemos observar en la Figura 2.1. Tomando una ventana pequeña, $h = 0.005$, podemos ver que la estimación es demasiado rugosa. Por el contrario, tomando una ventana grande, $h = 0.5$, la estimación resultante es demasiado plana.

De este modo, el gráfico (a) de la Figura 2.1 representa una estimación infrasuavizada, mientras que el gráfico (b) representa una estimación sobresuavizada. Así, hemos ilustrado para estos datos el enorme efecto que tiene la selección del parámetro ventana en el ajuste resultante, motivando así la necesidad de elegir h de alguna forma óptima.



(a) Estimación para $h = 0.005$.

(b) Estimación para $h = 0.5$.

Figura 2.1: Estimación lineal local con dos distintos parámetros de suavizado para los datos simulados a partir del modelo (2.1). En ambos gráficos la línea negra representa la función de regresión teórica del modelo.

Nuestro objetivo es encontrar un parámetro de suavizado h de forma que $\widehat{m}_h$ resulte óptimo en la estimación de m. Con la finalidad de comparar los diversos estimadores, es necesario disponer de criterios de error apropiados para medir la bondad de ajuste de los mismos.

Uno de los criterios de error usados más frecuentemente es el **Error Cuadrático Medio** (conocido habitualmente por sus siglas en inglés, MSE). Dado $\widehat{\theta}$ el estimador de un determinado parámetro θ, se define el error cuadrático medio del estimador $\widehat{\theta}$ como:

$$\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right] = \mathrm{Var}(\widehat{\theta}) + \left[\mathrm{Sesgo}(\widehat{\theta})\right]^2.$$

En el caso particular de $\widehat{m}_h(x)$, el estimador de la función $m(x)$ en un punto $x \in \mathbb{R}$ tenemos que:

$$\mathrm{MSE}\left(\widehat{m}_h(x)\right) = \mathbb{E}\left[(\widehat{m}_h(x) - m(x))^2\right]. \tag{2.2}$$

Notamos que se trata de un criterio de error local, puesto que depende tanto de x como de la ventana h. Así, el MSE será empleado como criterio de error cuando el objetivo principal sea estimar la función m en un punto x.

Por lo general, lo más deseable será estimar m en toda la recta real. Para evitar la dependencia del punto x, introducimos el **Error Cuadrático Integrado** (conocido habitualmente por sus

siglas en inglés, ISE), que se define como:

$$\text{ISE}(\widehat{m}_h) = \int \left[\widehat{m}_h(x) - m(x)\right]^2 w(x)dx,$$

donde $w(x) \geq 0$ es una función de peso. El ISE es un criterio de error global que no depende del punto en el que se evalúa el estimador, pero sí que depende de la muestra.

Con la finalidad de suprimir la aleatoriedad procedente de cada muestra individual, definimos el **Error Cuadrático Medio Integrado** (conocido habitualmente por sus siglas en inglés, MISE) de la siguiente forma:

$$\text{MISE}(\widehat{m}_h) = \mathbb{E}\left[\text{ISE}(\widehat{m}_h)\right] = \int \text{MSE}(\widehat{m}_h(x))w(x)dx, \tag{2.3}$$

que es un criterio de error global que no depende de la muestra empleada.

De este modo, tomaremos el MISE como criterio de error a minimizar para obtener un parámetro ventana óptimo desde el punto de vista teórico. Sin embargo, la expresión dada en (2.3) depende del parámetro ventana de una forma compleja, dificultando la interpretación de la influencia de este sobre la bondad del estimador. Es por esto que resulta de interés obtener una aproximación asintótica del MISE. La expresión asíntótica del MISE viene dada por:

$$\text{AMISE}(\widehat{m}_h) = \frac{1}{nh}R(K)\int v(x)\frac{w(x)}{f(x)}dx + \frac{1}{4}h^4\mu_2(K)^2\int m''(x)^2w(x)dx. \tag{2.4}$$

donde $\mu_j(K) = \int z^j K(z)dz$ y $R(K) = \int K(x)^2dx$, con $K \in L_2$, como puede consultarse en Fan y Gijbels (1996). Minimizando esta expresión obtenemos un parámetro ventana asintóticamente óptimo, que viene dado por:

$$h_{opt} = \left[\frac{R(K)\int v(x)\frac{w(x)}{f(x)}dx}{n\mu_2(K)^2\int m^{(2)}(x)^2w(x)dx}\right]^{-\frac{1}{5}}. \tag{2.5}$$

Este parámetro de suavizado depende de cantidades desconocidas como la densidad de la variable explicativa f, la varianza condicional σ^2 y la segunda derivada de la función de regresión $m''(x)$, que nos permite conocer la curvatura de la función de regresión. Por lo tanto, será necesario presentar estimadores de dichas cantidades para poder obtener selectores del parámetro de suavizado. En este punto, aunque no entremos más en detalle, podemos destacar los selectores *plug-in* que han sido propuestos por Ruppert y cols. (1995), o la regla del pulgar propuesta por Silverman (1986), cuyo objetivo es estimar (2.5) a través de estimaciones no paramétricas o paramétricas de las cantidades desconocidas. Por otra parte, también podemos considerar el método de validación cruzada, que consiste en tomar h tal que minimice la siguiente función:

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}_h^{-i}(X_i))^2,$$

donde $\widehat{m}_h^{-i}$ es el estimador lineal local donde la observación (X_i, Y_i) no ha sido tenida en cuenta para realizar el ajuste.

Ejemplo 2.2. Retomemos ahora el Ejemplo 2.1. En la Figura 2.1 representamos el ajuste lineal local para dos parámetros de suavizado distintos que seleccionamos arbitrariamente. Ahora, ajustaremos de nuevo un modelo de regresión lineal local, pero esta vez saleccionando h mediante algún procedimiento de optimización.

El comando `dpill` de  emplea la metodología *plug-in* para la selección de un parámetro ventana óptimo propuesto por Ruppert y cols. (1995) para la estimación lineal local, y este nos devuelve el valor $h = 0.043$. Tomando este parámetro de suavizado ajustamos un modelo de regresión lineal local, como podemos ver en la Figura 2.2. Observamos que en este caso la estimación se aproxima considerablemente a la función de regresión teórica del modelo (2.1) que representamos con la línea negra.

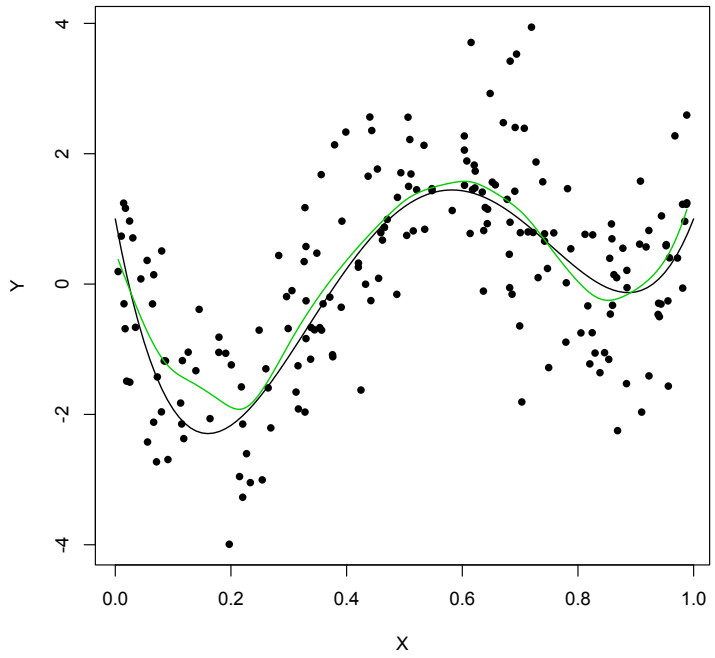


Figura 2.2: Ajuste de regresión lineal local usando el selector propuesto por Ruppert y cols. (1995), representado junto con la verdadera función de regresión del modelo (2.1).

Capítulo 3

Modelos de regresión aditivos

Los modelos de regresión aditivos, que fueron introducidos por Stone (1985), como ya hemos visto en la Sección 1.3, son modelos de la forma:

$$Y = \beta_0 + \sum_{j=1}^{p} m_j(X_j) + \varepsilon, \tag{3.1}$$

donde β_0 es el intercepto del modelo y las funciones m_i son desconocidas aunque se suponen funciones suaves, es decir, que cumplen unas ciertas condiciones de regularidad.

Los modelos de regresión aditivos son un caso particular de los **modelos aditivos generalizados** (habitualmente denotados por sus siglas en inglés GAM), propuestos por Hastie y Tibshirani (1986). Un modelo aditivo generalizado toma la forma:

$$g(\mathbb{E}[Y|X_1, \ldots, X_p]) = \beta_0 + \sum_{j=1}^{p} m_j(X_j),$$

donde a la función g se le conoce como función de enlace o función *link* y es una función conocida.

Como podemos observar, los modelos aditivos se corresponden con el caso en el que g es la función identidad. El objeto de estudio de este trabajo son los modelos aditivos, por lo que nos limitaremos al desarrollo de los mismos. Para ello, tomaremos como referencia fundamental Wood (2017), aunque esta referencia también se puede consultar si se desea profundizar en el estudio de los modelos GAM en general.

Acerca del modelo aditivo dado en (3.1), cabe destacar que se toma como hipótesis la aditividad de los efectos de las variables explicativas. Además, el hecho de que el modelo ahora incluya p funciones $m_1, \ldots, m_p$ introduce un problema de **identificabilidad**. Este puede ilustrarse del siguiente modo: si le sumamos una constante a m_1 y le restamos esa misma constante a m_2, las predicciones del modelo no sufrirán ningún cambio. En consecuencia, se deben imponer

restricciones de identificabilidad sobre el modelo antes de realizar el ajuste. Habitualmente, se impondrá como restricción que la media del efecto de cada variable sea 0, es decir:

$$\mathbb{E}[m_j(X_j)] = 0, \quad \text{para todo } j \in \{1, \dots, p\}. \tag{3.2}$$

La estimación de cada una de las funciones $m_1, \dots, m_p$ del modelo aditivo puede llevarse a cabo mediante técnicas no paramétricas como la regresión constante local o lineal local, que estudiamos en el Capítulo 2. De tomas formas, necesitaremos a mayores un procedimiento que "equilibre" las estimaciones de cada uno de los efectos. En esta línea, el **algorimo de backfitting** es un algoritmo general que permite el ajuste de un modelo aditivo empleando cualquier método de estimación. Se trata de un procedimiento de ajuste iterativo, y se formula como sigue:

1. *Inicializar los parámetros*: $\widehat{\beta}_0 = \overline{Y}$ y $\widehat{m}_j = \widehat{m}_j^0$, para todo $j \in \{1, \dots, p\}$.
2. *Iterar*: para $j = 1, \dots, p$:
$$\widehat{m}_j = S_j \left(Y - \widehat{\beta}_0 - \sum_{k \neq j} \widehat{m}_k(X_k) | X_j \right).$$
3. El proceso termina cuando las estimaciones de las funciones, $\widehat{m}_j$, converjan.

Denotamos por S_j al operador de suavizado de la respuesta respecto a la covariable X_j. Podemos tomar como operador cualquier técnica de estimación para las funciones m_j. Obviamente la complejidad de este tipo de procedimientos reside en la elección de los correspondientes parámetros de suavizado.

Si se desea profundizar en la estimación de un modelo aditivo mediante técnicas como la regresión lineal local a través de un algoritmo de backfitting, véase Hastie y Tibshirani (1990). Sin embargo, a la hora de estimar un modelo aditivo, la técnica empleada con más frecuencia se basa en la representación de cada función m_i en una base de funciones, que habitualmente estará conformada por *splines*. La representación de las funciones m_i en una base de funciones es lo que nos permitirá estimarlas mediante métodos de regresión penalizada y sacar así partido del método de mínimos cuadrados presentado en el Capítulo 1.

3.1. Caso particular: variable explicativa univariante

Para estudiar la representación y estimación de las funciones que componen el modelo aditivo, consideraremos primero el caso en el que tenemos una única variable explicativa. Es decir, consideraremos un modelo de la forma:

$$Y = m(X) + \varepsilon, \tag{3.3}$$

donde m es una función desconocida. Por simplicidad, vamos a suponer que X toma valores en el intervalo [0,1].

Podemos tomar una base de funciones $b_1, \ldots, b_k$ de un espacio de funciones que contenga a m, de modo que la función m se pueda expresar de la forma:

$$m(X) = \sum_{i=1}^{k} b_i(X)\beta_i, \tag{3.4}$$

donde $\beta_1, \ldots, \beta_k$ son parámetros desconocidos.

Supongamos que conocemos $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ una muestra aleatoria simple de las variables X e Y. Sustituyendo la expresión (3.4) en el modelo (3.3), observamos que obtenemos un modelo lineal, dado por

$$\mathbb{Y} = \mathbb{X}\beta + \varepsilon, \tag{3.5}$$

donde $\mathbb{X}$ es una matriz de dimensión $n \times (k+1)$, cuyo elemento i, j viene dado por $b_j(X_i)$. Nótese que en la expresión (3.5), $\mathbb{Y} = (Y_1, \ldots, Y_n)'$ y ε denotan vectores de dimensión n y la primera columna de $\mathbb{X}$ está comuesta por unos.

De este modo, hemos expresado el modelo (3.3) en la forma de un modelo lineal. Esto es lo que nos permite ajustarlo mediante el método de mínimos cuadrados, es decir, minimizando $||\mathbb{Y} - \mathbb{X}\beta||^2$.

Ejemplo 3.1. Si asumimos que m es un polinomio de orden q, el espacio de polinomios de orden menor o igual que q contendrá a m. Así, si consideramos la base de este espacio:

$$b_1(X) = 1,\ b_2(X) = X,\ b_3(X) = X^2,\ b_4(X) = X^3, \ldots, b_{q+1}(X) = X^q,$$

podríamos expresar m de la forma

$$m(X) = \beta_1 + \beta_2 X + \beta_3 X^2 + \beta_4 X^3 + \cdots + \beta_{q+1} X^q$$

y el modelo dado en (3.3) podría escribirse como sigue:

$$Y = \beta_1 + \beta_2 X + \beta_3 X^2 + \beta_4 X^3 + \cdots + \beta_{q+1} X^q + \varepsilon.$$

Obviamente, a la vista de este ejemplo, es claro que si la dimensión de la base es menor que $q+1$ no seremos capaces de reproducir bien el comportamiento de nuestro modelo.

Las bases polinómicas no dan muy buenos resultados en la práctica, por lo que representaremos ahora la función m en una base de funciones lineales a trozos. Posteriormente veremos que los resultados dados por este tipo de bases pueden mejorarse empleando bases que estén formadas por *splines*. Sin embargo, estudiaremos primero la representación de m en una base lineal a trozos a modo de ilustración, puesto que su interpretación resulta más intuitiva.

3.1.1. Representación de una función en una base lineal a trozos

Para ilustrar este tipo de ajustes no paramétricos, lo haremos, como ya hemos comentado, usando una base de funciones lineales a trozos.

Una base de **funciones lineales a trozos** está completamente determinada por las localizaciones de las discontinuidades de la derivada de dichas funciones, es decir, los puntos donde los trozos lineales se unen. Estos puntos se denominan **nodos** y los denotaremos por $\{z_j : j = 1, \ldots, k\}$ con $z_{j-1} < z_j$ para todo $j \in \{1, \ldots, k\}$. Dados unos determinados nodos, podemos considerar la siguiente base para las funciones lineales a trozos:

$$
\begin{aligned}
b_1(x) &= \begin{cases} \frac{z_2 - x}{z_2 - z_1} & \text{si } x < z_2, \\ 0 & \text{en otro caso.} \end{cases} \\
b_j(x) &= \begin{cases} \frac{x - z_{j-1}}{z_j - z_{j-1}} & \text{si } z_{j-1} < x \leq z_j, \\ \frac{z_{j+1} - x}{z_{j+1} - z_j} & \text{si } z_j < x < z_{j+1}, \\ 0 & \text{en otro caso,} \end{cases} \qquad \text{para } j = 2, \ldots, k-1. \\
b_k(x) &= \begin{cases} \frac{x - z_{k-1}}{z_k - z_{k-1}} & \text{si } x > z_{k-1}, \\ 0 & \text{en otro caso.} \end{cases}
\end{aligned}
\tag{3.6}
$$

A modo de ejemplo, tomamos como nodos 6 puntos equiespaciados en el intervalo $[0, 1]$. Entonces, las funciones que conforman la base dada en (3.6) determinada por estos nodos son las representadas en la Figura 3.1. Observamos que cada función b_i de la base toma el valor 1 en el respectivo nodo z_i y 0 en el resto.

Ejemplo 3.2. Consideremos nuevamente los datos simulados a partir del modelo univariante (2.1), que recordamos que venía dado por:

$$Y = m(X) + \varepsilon = 1 - 48X + 218X^2 - 315X^3 + 145X^4 + \varepsilon.$$

Ajustaremos varios modelos de regresión empleando bases de funciones lineales a trozos para estos datos. Para el ajuste de cada uno de estos modelos es necesaria una previa selección del conjunto de nodos. En este caso, tomaremos respectivamente $k = 3$, $k = 10$ y $k = 40$ puntos equiespaciados en el intervalo $[0, 1]$, por lo que estaremos empleando bases de 3 tamaños distintos en cada uno de los ajustes. Los ajustes resultantes se pueden ver representados en la Figura 3.2, junto con la verdadera función de regresión asociada al modelo (2.1). Destacamos que la elección de la dimensiones de las bases ha sido completamente arbitraria.

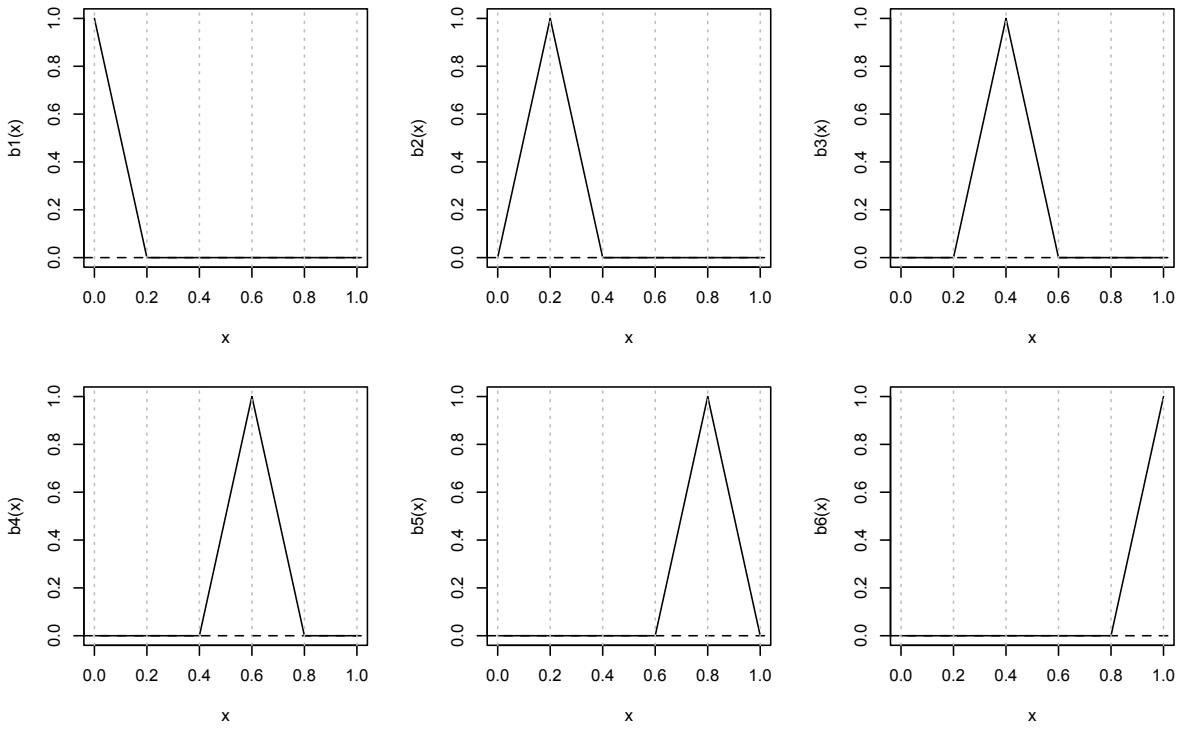


Figura 3.1: Base de funciones lineales a trozos determinada por 6 nodos equiespaciados en el intervalo $[0, 1]$.

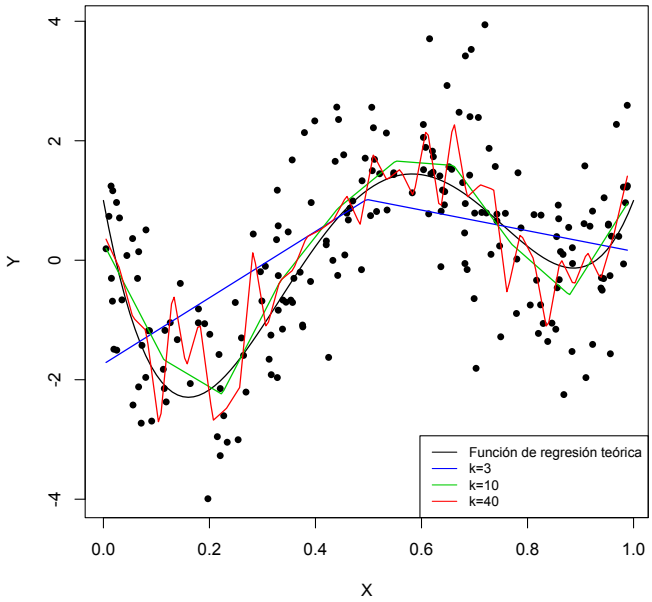


Figura 3.2: Ajustes de regresión usando una base de funciones lineales a trozos para distintos tamaños de base k para los datos simulados a partir del modelo (2.1).

Observamos que para llevar a cabo el ajuste del modelo es necesaria una previa elección de los nodos, así como de la dimensión de la base, k, que es el parámetro que determina el grado de suavidad que tendrá la estimación de la función m. A la vista de la Figura 3.2 se pone de manifiesto que según aumentamos el tamaño de la base la estimación se vuelve más rugosa. Es por esto que conviene seleccionar k de alguna forma óptima, aunque no se trate de una tarea sencilla.

Así, en lugar de alterar la dimensión de la base para controlar el grado de suavidad de la estimación de m empleamos la siguiente alternativa: mantenemos constante la dimensión de la base, a un valor mayor del que se espera que sea necesario, e introducimos un término de penalización de rugosidad al ajuste por mínimos cuadrados, es decir, una penalización al hecho de incluir muchos parámetros en el modelo estimado.

De este modo, en lugar de ajustar el modelo minimizando $||\mathbb{Y} - \mathbb{X}\beta||^2$, lo haremos mediante un método de mínimos cuadrados penalizado, es decir, minimizando la expresión:

$$||\mathbb{Y} - \mathbb{X}\beta||^2 + \lambda \sum_{j=2}^{k-1} \left[m(z_{j-1}) - 2m(z_j) + m(z_{j+1})\right]^2, \tag{3.7}$$

donde el término del sumatorio penaliza los modelos que sean demasiado rugosos. Para la estimación mediante otros tipos de *splines* más complejos, este término se sustituirá por una penalización del tipo $\int [m''(x)]^2 dx$. El parámetro de suavizado, que denotamos por λ, es el que controlará el equilibrio entre la suavidad de la función estimada y que interpole los datos de la muestra. Para $\lambda = 0$, tenemos una estimación mediante *splines* sin penalización, y según λ tiende a infinito, la estimación se convierte en una línea recta.

En general, dada la formulación del modelo presentada en (3.5), el término de la penalización podrá escribirse como una forma cuadrática en β. En concreto para esta base lineal a trozos, los coeficientes β_j se corresponden con los valores de m en los nodos, es decir, $\beta_j = m(z_j)$, por lo que la penalización se puede escribir del siguiente modo:

$$\sum_{j=2}^{k-1} \left[\beta_{j-1} - 2\beta_j + \beta_{j+1}\right]^2 = \beta' D' D \beta = \beta S \beta, \tag{3.8}$$

donde

$$D = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \dots \\ 0 & 1 & -2 & 1 & 0 & \dots \\ 0 & 0 & 1 & -2 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

En consecuencia, el ajuste del modelo será el resultado de minimizar:

$$||\mathbb{Y} - \mathbb{X}\beta||^2 + \lambda \beta' S \beta. \tag{3.9}$$

Es decir, hemos conseguido reescribir nuestro problema de estimación no paramétrica como un problema de estimación por mínimos cuadrados ponderados, lo que supone un importante avance computacional a la hora de implementar este tipo de modelos. Este hecho es el que motiva que los modelos aditivos ajustados mediante *splines* sean mucho más usados en la práctica que aquellos que están basados en ideas de suavización tipo núcleo como las presentadas en el Capítulo 2.

3.1.2. Estimación del vector de parámetros

Es sencillo ver que la solución de (3.9), es decir, el estimador por mínimos cuadrados penalizados de β, de manera análoga a lo viso en el Capítulo 1, está dado por:

$$\widehat{\beta} = (\mathbb{X}'\mathbb{X} + \lambda S)^{-1}\mathbb{X}'\mathbb{Y}.$$

Además, para mejorar el coste computacional del ajuste planteado podemos expresar el problema de minimización (3.9) de la siguiente forma alternativa:

$$||\mathbb{Y} - \mathbb{X}\beta||^2 + h\beta' S\beta = \left\| \begin{bmatrix} \mathbb{Y} \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbb{X} \\ \sqrt{h}D \end{bmatrix} \beta \right\|^2,$$

que se conoce como método de mínimos cuadrados aumentado, y que tiene asociado un menor coste computacional.

3.1.3. Selección del parámetro de suavizado

Tal y como hemos planteado el ajuste, podemos observar que para estimar el modelo (3.3), habría que elegir el tamaño de la base, k, las localizaciones de los nodos, z_j, y un valor para el parámetro de suavizado λ. Sin embargo, si tomamos k mayor de lo que cabe esperar que sea necesario para representar m, como ya hemos mencionado anteriormente, ni el valor exacto de k ni la localización precisa de los nodos van a tener una gran influencia en el ajuste. Es la elección del parámetro de suavizado λ la que resulta crucial en la determinación de la bondad del ajuste asociado a este tipo de modelos.

El problema de estimar el grado de suavidad del modelo se reduce a estimar el parámetro de suavizado λ. De manera análoga a los que vimos en la Sección 2.2.2, el parámetro de suavizado λ tiene una gran influencia en el ajuste del modelo: si tomamos un λ demasiado alto obtendremos una estimación sobresuavizada, y si tomamos un λ demasiado pequeño la estimación estará infrasuavizada.

Denotamos por $\widehat{m}_\lambda$ a la estimación de m obtenida mediante regresión por *splines* con parámetro de suavizado asociado λ. Para poder seleccionar el parámetro de suavizado, la idea más

intuitiva será considerar λ tal que se minimice la expresión:

$$M = \int (m(x) - \widehat{m}_\lambda(x))^2 dx,$$

es decir, que el error cuadrático medio de la estimación con respecto al valor real sea lo más pequeño posible.

Evidentemente, la función m es desconocida, por lo que no es posible calcular M directamente. Así, definimos la **validación cruzada ordinaria** (habitualmente conocida por sus siglas en inglés, OCV) como sigue:

$$\text{OCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{m}_\lambda^{-i}(X_i)\right)^2, \tag{3.10}$$

donde $\widehat{m}_\lambda^{-i}$ es el estimador de m donde la observación (X_i, Y_i) no ha sido tenida en cuenta para realizar el ajuste. Este estimador se conoce habitualmente como *leave-one-out* y ya lo hemos presentado anteriormente en el Capítulo 2 en el contexto de regresión tipo núcleo.

Podemos tomar λ tal que minimice (3.10), lo que se conoce como método de validación cruzada ordinaria. El cálculo de OCV, sin embargo, es computacionalmente muy costoso puesto que debemos evaluar (3.10) en una rejilla de posibles valores del parámetro de suavizado.

Se verifica que:

$$\text{OCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{m}_\lambda(X_i)}{1 - H_{ii}}\right)^2,$$

donde $H = \mathbb{X}(\mathbb{X}'\mathbb{X} + \lambda S)^{-1}\mathbb{X}'$ es la matriz *hat* del modelo. Sustituyendo los pesos, H_{ii}, por el peso medio, $tr(H)/n$, obtenemos la denominada **validación cruzada generalizada** (habitualmente conocida por sus siglas en inglés, GCV), que viene dada por:

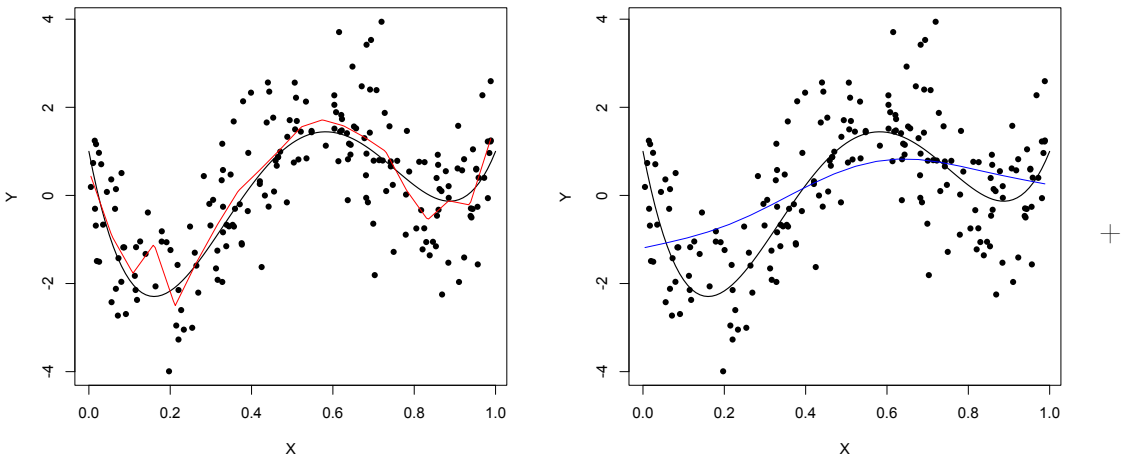
$$\text{GCV}(\lambda) = \frac{n\sum_{i=1}^{n}\left(Y_i - \widehat{m}_\lambda(X_i)\right)^2}{\left(n - tr(H)\right)^2}. \tag{3.11}$$

En general, es preferible el uso de la GCV como criterio a minimizar para la elección de λ debido a sus mejores propiedades computacionales.

Ejemplo 3.3. Retomamos los datos simulados a partir del modelo univariante (2.1). Ya presentamos en la Figura 3.2 varias estimaciones obtenidas empleando una base lineal a trozos y ajustando el modelo mediante mínimos cuadrados. Veamos ahora la estimación de estos datos resultante de minimizar la expresión dada en (3.9), es decir, el método de mínimos cuadrados que introduce una penalización de suavidad.

Recordamos que en este caso ya no es el tamaño de la base el que determina el grado de suavidad, sino que este se mantiene fijo y se debe seleccionar un determinado parámetro de suavizado λ. Como podemos apreciar en la Figura 3.3, un parámetro de suavizado demasiado

pequeño producirá una estimación infrasuavizada (parte (a) de la Figura 3.3), mientras que un parámetro de suavizado demasiado grande producirá una estimación sobresuavizada (parte (b) de la Figura 3.3).



(a) Estimación para $\lambda = 0.1$.

(b) Estimación para $\lambda = 1000$.

Figura 3.3: Ajuste usando una base de funciones lineales a trozos penalizada para los datos simulados del modelo (2.1), tomando dos parámetros de suavizado λ distintos.

Para seleccionar el parámetro de suavizado de forma óptima, procedemos buscando λ tal que minimice la GCV dada en (3.11). En la Figura 3.4 podemos ver representado el valor de GCV(λ) frente al logaritmo de λ para los datos simulados del modelo (2.1), y vemos que el mínimo se obtiene para $\lambda = 127.547$.

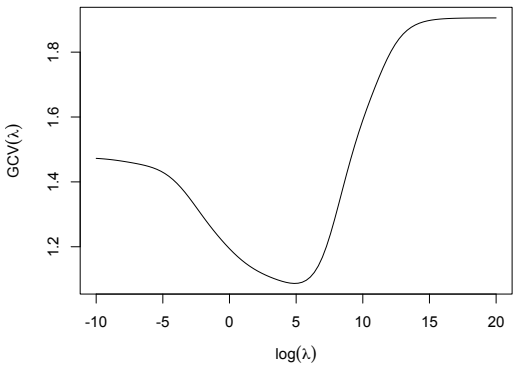


Figura 3.4: Representación de la función GCV frente al logaritmo de λ para los datos simulados del modelo (2.1)

Así, tomando este parámetro de suavizado, que consideramos óptimo, y ajustando el modelo por mínimos cuadrados penalizados obtendríamos la estimación representada en la Figura 3.5 para los datos del modelo (2.1). En la Figura 3.5 también representamos mediante la línea negra la función teórica del modelo. De este modo, podemos observar que tomando un λ óptimo mediante el criterio GCV, la estimación resultante es aceptable.

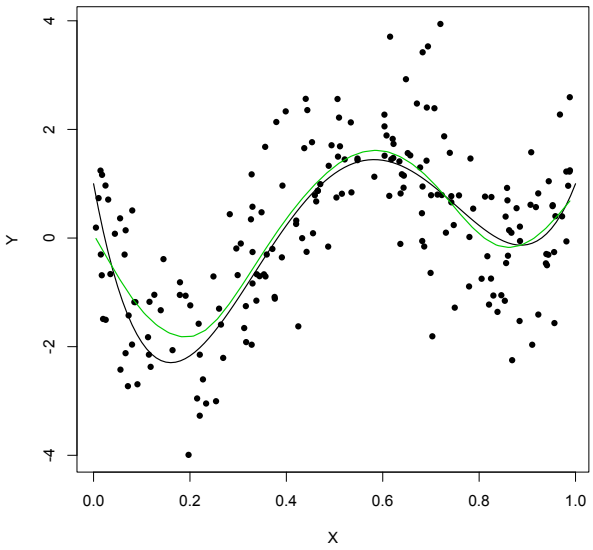


Figura 3.5: Ajuste del modelo usando una base de funciones lineales a trozos penalizada para los datos simulados del modelo (2.1), tomando como parámetro de suavizado λ aquel que minimiza la función GCV.

3.2. Bases de funciones *spline*

El suavizador lineal a trozos considerado en la Sección 3.1 es un ejemplo sencillo y razonable para representar los efectos de las variables explicativas sobre la variable respuesta, pero los resultados dados por esta base son muy mejorables. Lo más habitual en este tipo de contextos es el empleo de bases de funciones *spline*, que reducen sustancialmente el error cometido para un determinado tamaño de la base debido a su flexibilidad frente a las funciones lineales a trozos consideradas anteriormente. En concreto, vamos a considerar una base formada por *splines* cúbicos, que definimos a continuación.

Definición 3.4. Un ***spline* cúbico** es una curva formada por secciones de polinomios cúbicos unidos de forma que la curva resultante sea continua hasta su segunda derivada. Los puntos de unión de las secciones son los que se denominan nodos del *spline*. Un *spline* cúbico se denomina **natural** si sus segundas derivadas en los nodos extremos son nulas.

Consideremos un conjunto de puntos $\{X_i, Y_i : i = 1, \ldots, n\}$, con $X_i < X_{i+1}$ para todo $i \in \{1, \ldots, n-1\}$. El *spline* cúbico natural interpolando estos puntos, que llamaremos $s(x)$, es una función formada por secciones de un polinomio de tercer orden, cada una definida en un intervalo $[X_i, X_{i+1}]$, y unidas de forma que s es de clase 2 con $s(X_i) = Y_i$ y $s''(X_1) = s''(X_n) = 0$. Es decir, es una función que satisface las siguientes condiciones:

1. $s(x) \in C^2[X_1, X_n], \forall x \in [X_1, X_n]$.
2. En cada intervalo $[X_i, X_{i+1}]$, con $i \in \{1 \ldots, n-1\}$, $s(x)$ es un polinomio de tercer orden.
3. $s(X_i) = Y_i$ con $i \in \{1 \ldots, n\}$.
4. $s''(X_1) = s''(X_n) = 0$.

En la Figura 3.6 podemos ver la representación del *spline* cúbico natural que interpola los 6 puntos destacados en el gráfico, que son $\{(0, 0.1), (0.2, 0.4), (0.4, 0.2), (0.6, 0.9), (0.8, 0.4), (1, 0.6)\}$. Es decir, está formado por 5 secciones de polinomios de tercer orden.

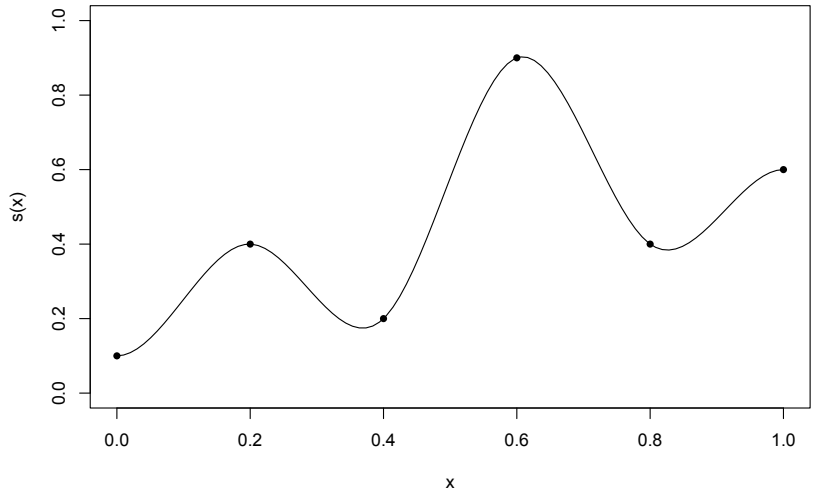


Figura 3.6: *Spline* cúbico natural interpolando los puntos $\{(0, 0.1), (0.2, 0.4), (0.4, 0.2), (0.6, 0.9), (0.8, 0.4), (1, 0.6)\}$.

Denotemos por $S_2[a, b]$ el espacio de funciones continuas en el intervalo $[a, b]$ con primera derivada absolutamente continua[1]. De entre todas las funciones g pertenecientes a $S_2[X_1, X_n]$ que interpolan los puntos $\{X_i, Y_i : i = 1, \ldots, n\}$, el *spline* cúbico $s(x)$ es el óptimo en el sentido de que minimiza:

$$\int_{X_1}^{X_n} g''(x)^2 dx.$$

[1]Decimos que una función g es absolutamente continua en un intervalo $[a, b]$ si existe una función integrable g' tal que $g(x) = g(a) + \int_a^x g'(t)dt$.

Podemos ver una demostración de este resultado en la página 16 de Green y Silverman (1994).

3.2.1. Estimación usando *splines* cúbicos

Por lo general, nuestro propósito es suavizar los datos y no interpolarlos. Por lo tanto, en lugar de establecer $s(X_i) = Y_i$, trataremos los $s(X_i)$ como los n parámetros libres del *spline* cúbico, y los estimaremos de forma que se minimice la siguiente expresión:

$$\sum_{i=1}^{n} [Y_i - s(X_i)]^2 + \lambda \int s''(x)^2 dx,$$

donde λ es un parámetro de suavizado. La función $s(x)$ resultante es lo que denominamos como ***spline* de suavizado**. De entre las funciones que pertenecen a $S_2[X_1, X_n]$, se puede comprobar de manera inmediata que s es la que minimiza:

$$\sum_{i=1}^{n} [Y_i - g(X_i)]^2 + \lambda \int g''(x)^2 dx. \tag{3.12}$$

Así, las bases de *splines* cúbicos surgen como el resultado de minimizar la expresión (3.12), por lo que se pueden considerar óptimas. Esto justifica que tomemos una base de *splines* cúbicos para estimar cada uno de los efectos que aglutina un modelo aditivo.

Podemos representar un *spline* cúbico mediante diversas bases equivalentes. En este caso presentaremos una base que parametriza el *spline* en términos de los valores que este toma en los nodos. Sea $s(x)$ un *spline* cúbico con nodos $z_1, \ldots, z_k$. Sean $\beta_j = s(z_j)$ y $\delta_j = s''(z_j)$, con $j \in \{1, \ldots, k\}$. Definimos:

$$a_j^-(x) = \frac{z_{j+1} - x}{h_j}, \qquad c_j^-(x) = \frac{1}{6}\left[\frac{(z_{j+1} - x)^3}{h_j} - h_j(z_{j+1} - x)\right],$$
$$a_j^+(x) = \frac{x - z_j}{h_j}, \qquad c_j^+(x) = \frac{1}{6}\left[\frac{(x - z_j)^3}{h_j} - h_j(x - z_j)\right],$$

donde $h_j = z_{j+1} - z_j$. De este modo, podemos escribir:

$$s(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\delta_j + c_j^+(x)\delta_{j+1}, \text{ si } z_j \leq x \leq z_{j+1}. \tag{3.13}$$

Por hipótesis, se tiene que s es de clase 2 en los nodos y $\delta_1 = \delta_k = 0$ (consideramos un *spline* cúbico natural). Así, se puede ver que:

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

donde $\delta^- = (\delta_2, \ldots, \delta_{k-1})'$ y donde los elementos no nulos de las matrices D y B son los siguientes:

$$\begin{aligned}
&D_{i,i} = \frac{1}{h_i} && D_{i,i+1} = -\frac{1}{h_i} - \frac{1}{h_{i+1}} && D_{i,i+2} = \frac{1}{h_{i+1}}, && i = 1, \ldots, k-2,\\
&B_{i,i} = \frac{h_i + h_{i+1}}{3}, && && && i = 1, \ldots, k-2,\\
&B_{i,i+1} = \frac{h_{i+1}}{6} && B_{i+1,i} = \frac{h_{i+1}}{6}, && && i = 1, \ldots, k-3.
\end{aligned}$$

Podemos escribir $\delta = F\beta$, donde

$$F = \begin{bmatrix} 0 \\ B^{-1}D \\ 0 \end{bmatrix}$$

de modo que la expresión (3.13) puede reescribirse exclusivamente en términos de los parámetros $\beta_1, \ldots, \beta_k$ como sigue:

$$s(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)F_j\beta + c_j^+(x)F_{j+1}\beta, \text{ si } z_j \leq x \leq z_{j+1}. \tag{3.14}$$

De esta forma, el *spline* s puede expresarse más convenientemente de la forma:

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

donde $b_i(x)$ son las funciones que forman la base, definidas implícitamente a partir de la expresión (3.14). A modo de ejemplo, en la Figura 3.7 ilustramos los elementos de una base de *splines* cúbicos naturales, obtenida tomando como nodos 6 puntos equiespaciados en el intervalo $[0, 1]$.

En cuanto a la estimación de los parámetros, conviene tener en cuenta que la matriz de penalización respectiva a esta base viene dada por $S = D'B^{-1}D$, ya que

$$\int_{z_1}^{z_k} s''(x)^2 dx = \beta' D' B^{-1} D\beta,$$

como se puede ver en la Sección 4.7 de Lancaster y Salkauskas (1986).

Además del *spline* cúbico, existen otras técnicas de suavizado de una función basadas en *splines*. Destacamos los ***splines* cúbicos cíclicos**, que son iguales que los anteriores, pero tomando como hipótesis que el punto inicial coincide con el final; y los **P-*splines***, que se desarrollan a partir de una base de B-*splines*, introduciendo una penalización. En la práctica estas bases son muy eficientes, pero tienen algunas desventajas: dependen de los nodos elegidos y sólo son útiles para suavizar modelos en los que haya una única variable explicativa. Una solución a estos problemas son los ***splines* de placa delgada** o TPS (*thin plate splines*) que además poseen

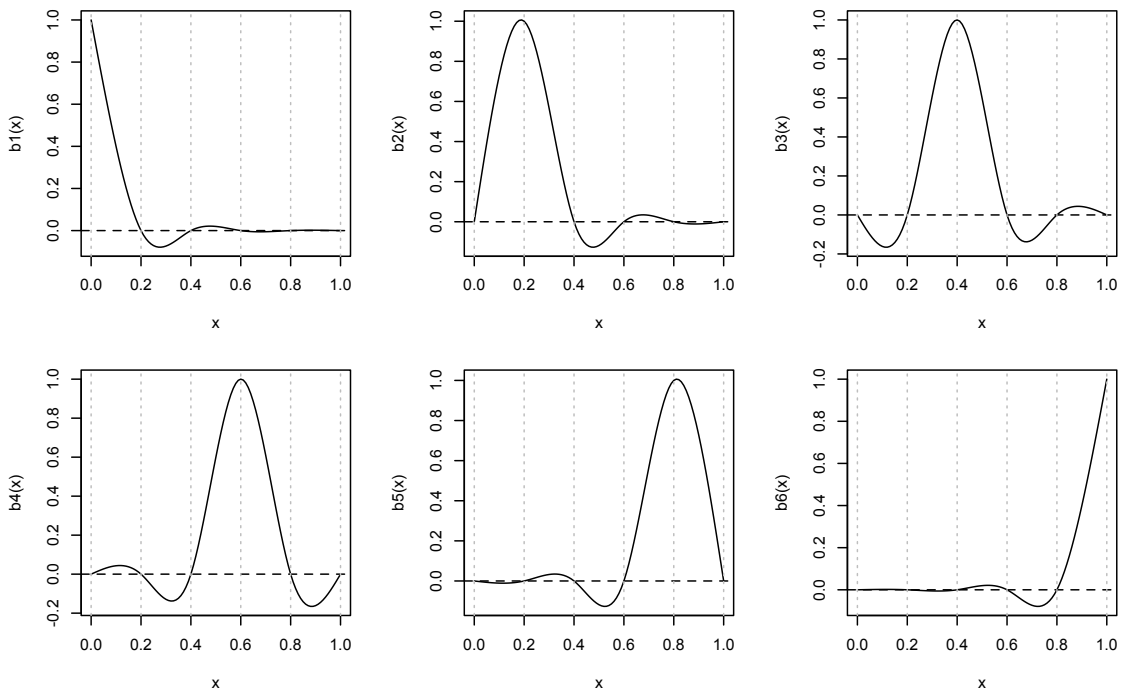


Figura 3.7: Base de *splines* cúbicos naturales determinada por 6 nodos equiespaciados en el intervalo $[0, 1]$.

diversas propiedades de optimalidad. Sin embargo, su elevado coste computacional para grandes conjuntos de datos es lo que en muchos casos nos lleva a utilizar alguno de los *splines* presentados anteriormente. Véase Wood (2017) si se desea profundizar en los distintos tipos de *splines* y sus propiedades como estimadores para una función suave.

Ejemplo 3.5. Retomamos una vez más los datos simulados a partir del modelo (2.1), cuya expresión estaba dada por:

$$Y = m(X) + \varepsilon = 1 - 48X + 218X^2 - 315X^3 + 145X^4 + \varepsilon.$$

Obtenemos para estos datos un ajuste de regresión empleando *splines* cúbicos, como hemos visto a lo largo de esta sección. En la Figura 3.8 representamos este ajuste junto con otras estimaciones obtenidas empleando los distintos tipos de *splines* que hemos mencionado anteriormente.

A la vista de la Figura 3.8 es evidente que la elección de la base de *splines* que empleamos para el ajuste no tiene demasiada relevancia en el desempeño del estimador. Como podemos ver, las diferencias entre las distintas estimaciones son mínimas, únicamente apreciables cerca de los valores extremos que toma la variable explicativa.

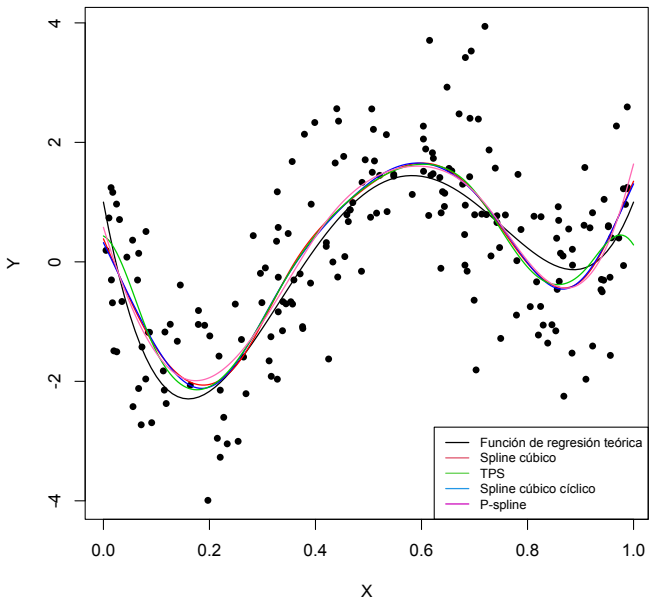


Figura 3.8: Ajuste no paramétrico basado en distintos tipos de *splines* de regresión para los datos simulados a partir del modelo (2.1).

3.3. El modelo aditivo

Retomamos ahora el modelo aditivo, cuya expresión recordamos que está dada por:

$$Y = \beta_0 + \sum_{j=1}^{p} m_j(X_j) + \varepsilon. \tag{3.15}$$

donde se supone que se cumple la condición de identificabilidad (3.2), es decir:

$$\mathbb{E}[m_j(X_j)] = 0, \quad \text{para todo } j \in \{1, \ldots, p\},$$

y además el error del modelo sigue una distribución Gaussiana.

Entonces, cada uno de los efectos de las variables explicativas involucradas en el modelo aditivo pueden ser representados empleando bases de funciones *splines*, estimadas mediante mínimos cuadrados penalizados y su grado de suavidad, es decir, el parámetro de suavizado λ, elegido por criterios de validación cruzada generalizada, empleando el mismo procedimiento que seguimos para el caso univariante.

De este modo, cada uno de los efectos m_j del modelo, con $j \in \{1, \ldots, p\}$, puede ser representado empleando una base de funciones *spline* definida a partir de los nodos $\{z_{l,j} : l = 1, \ldots, k_j\}$. Así, tendremos la siguiente representación:

$$m_j(X_j) = \sum_{i=1}^{k_j} b_i(X_j)\beta_{i,j} \tag{3.16}$$

donde $\beta_{1,j}, \ldots, \beta_{k_j,j}$ son los coeficientes desconocidos y $b_1(X_j), \ldots, b_{k_j}(X_j)$ son las funciones *spline* de la base asociadas a X_j. Si empleamos *splines* cúbicos, como vimos en la Sección 3.2, estas funciones serán las definidas implícitamente a partir de la expresión (3.14).

Para solucionar el problema de la identificabilidad, como ya hemos comentado anteriormente, impondremos para todo $j \in \{1, \ldots, p\}$:

$$\sum_{i=1}^{n} m_j(X_{i,j}) = 0.$$

La forma más simple de hacer efectivas las restricciones de identificabilidad es considerar $\beta_{1,j} = 0$ para todo $j = 1, \ldots, p$. De esta forma, para cada $j \in \{1, \ldots, p\}$, denotamos por $\mathbb{X}_j$ a la matriz de dimensiones $n \times k_j$ cuyo elemento i, l viene dado por $b_l(X_{i,j})$, con $i = 1, \ldots, n$ y $l = 2 \ldots, k_j$ (la columna correspondiente a b_1 se completa con ceros), y denotamos $\beta_j' = (0, \beta_{2,j}, \ldots, \beta_{k_j,j})$.

Sustituyendo la expresión (3.16) en el modelo (3.15) logramos reescribirlo como un modelo lineal de la forma:

$$\mathbb{Y} = \mathbb{X}\beta + \varepsilon,$$

donde $\mathbb{X} = [1, \mathbb{X}_1, \ldots, \mathbb{X}_p]$ y el vector de parámetros es $\beta' = (\beta_0, \beta_1', \ldots, \beta_p')$.

De forma análoga a lo visto en la Sección 3.1, ajustaremos el modelo mediante mínimos cuadrados ponderados. De igual modo que vimos en (3.8), la penalización de suavidad asociada a cada efecto puede expresarse en la forma matricial $\beta_j' \bar{S}_j \beta_j = \beta_j' D_j' D_j \beta_j$. Además, esta penalización podrá expresarse más convenientemente como sigue:

$$\beta' S_j \beta = (\beta_0, \beta_1', \ldots, \beta_p') \begin{bmatrix} 0 & \ldots & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ 0 & \ldots & \bar{S}_i & \ldots & 0 \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \beta_j' \bar{S}_j \beta_j.$$

Ahora, la estimación del modelo (3.15) se obtiene mediante el método de mínimos cuadrados penalizados, es decir, minimizando la siguiente expresión:

$$||\mathbb{Y} - \mathbb{X}\beta||^2 + \lambda_1 \beta' S_1 \beta + \cdots + \lambda_p \beta' S_p \beta,$$

donde cada λ_j es el parámetro de suavizado asociado al efecto de la variable X_j, con $j = 1, \ldots, p$. Podemos reescribir la función a minimizar como:

$$||\mathbb{Y} - \mathbb{X}\beta||^2 + \lambda_1 \beta' S_1 \beta + \cdots + \lambda_p \beta' S_p \beta = \left\| \begin{bmatrix} \mathbb{Y} \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbb{X} \\ B \end{bmatrix} \beta \right\|^2,$$

donde

$$B = \begin{bmatrix} 0 & \sqrt{\lambda_1}D_1 & 0 & \dots & 0 \\ 0 & 0 & \sqrt{\lambda_2}D_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sqrt{\lambda_p}D_p \end{bmatrix}$$

o cualquier otra matriz tal que $B'B = \lambda_1 S_1 + \cdots + \lambda_p S_p$. Así, tenemos un método de mínimos cuadrados no penalizado para una versión aumentada del modelo, por lo que el modelo aditivo dado en (3.15) puede ser ajustado mediante técnicas de mínimos cuadrados combinadas con GCV para seleccionar los parámetros de suavizado $\lambda_1, \dots, \lambda_p$.

Ejemplo 3.6. Consideramos el modelo aditivo (1.5) formulado al final del Capítulo 1. Recordamos que este tomaba la forma siguiente:

$$Y = m(X_1, X_2, X_3, X_4) + \varepsilon = m_1(X_1) + m_2(X_2) + m_3(X_3) + m_4(X_4) + \varepsilon,$$

donde $\varepsilon \in N(0, 1)$ y los efectos de cada variable explicativa sobre la variable respuesta venían dados por:

$$\begin{aligned} m_1(X_1) =& 5X_1, \\ m_2(X_2) =& 1 - 48X_2 + 218X_2^2 - 315X_2^3 + 145X_2^4, \\ m_3(X_3) =& \sin(5\pi X_3), \\ m_4(X_4) =& 10(X_4^4 + X_4^2 - X_4). \end{aligned}$$

En el Ejemplo 1.1 generamos unos datos simulados a partir de este modelo. A continuación emplearemos estos datos para el ajuste de un modelo aditivo.

Para la implementación del ajuste de los modelos aditivos emplearemos el paquete `mgcv` de  (véase Wood (2023)). Este permite ajustar modelos aditivos generalizados mediante una gran variedad de *splines*, y por defecto realiza una selección automática de los parámetros de suavizado mediante algún criterio de optimización. También existe el paquete `gam` (véase Hastie (2023)), que además de *splines* de suavizado admite otros métodos como regresión polinómica local. Sin embargo, no dispone de un criterio de elección óptima de los parámetros de suavizado, por lo que estos deben ser especificados por el usuario. Esto justifica que el uso del paquete `mgcv` sea más recomendable en la mayoría de los casos.

Ajustamos un modelo aditivo empleando *splines* cúbicos para representar los datos simulados. Tomamos como dimensión de la base de *splines* $k = 10$, pues podemos comprobar que si aumentamos este valor no se produce ningún cambio notable en los ajustes producidos por el modelo. En consecuencia, nos basta con tomar $k = 10$. El código de  necesario para llevar a cabo el ajuste es el siguiente:

```
> library(mgcv)
> ma <- gam(y ~ s(x1,k=10,bs="cr") + s(x2,k=10,bs="cr") + s(x3,k=10,bs="cr")
  + s(x4,k=10,bs="cr"))
> summary(ma)
Family: gaussian
Link function: identity

Formula:
y ~ s(x1, k = 10, bs = "cr") + s(x2, k = 10, bs = "cr") + s(x3,
k = 10, bs = "cr") + s(x4, k = 10, bs = "cr")

Parametric coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.34549    0.07551   44.31   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
edf Ref.df      F p-value
s(x1) 1.999  2.508 139.67  <2e-16 ***
s(x2) 6.806  7.866  17.66  <2e-16 ***
s(x3) 7.674  8.520  10.48  <2e-16 ***
s(x4) 4.729  5.740 280.44  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.925   Deviance explained = 93.3%
GCV = 1.2826  Scale est. = 1.1402    n = 200
```

Observamos que al aplicar el comando `summary` al modelo `ma` obtenemos el porcentaje de variabilidad explicada por el modelo (`Deviance explained = 93.3%`). Este porcentaje es una medida de la bondad del ajuste. En este caso, el porcentaje de variabilidad explicada es muy alto, por lo que podemos concluir que el ajuste obtenido es bueno.

Notamos que no es necesaria la especificación de la dimensión de las bases de *splines* para la estimación de cada uno de los efectos del modelo, sino que podríamos ajustar un modelo aditivo de la siguiente forma:

```
> ma0 <- gam(y ~ s(x1) + s(x2) + s(x3) + s(x4))
```

en cuyo caso la dimensión de cada base de *splines* sería seleccionada automáticamente y por defecto estarían formadas por TPS. Podemos comprobar que los resultados dados por ambos modelos son prácticamente iguales.

Aplicando el comando `plot` al modelo aditivo `ma`, obtenemos una representación de las estimaciones de los efectos de cada una de las variables X_1, X_2, X_3 y X_4 sobre la variable respuesta, que se corresponden con los 4 gráficos situados a la izquierda en la Figura 3.9. Además, en estas representaciones también se representan intervalos de confianza del 95 % para dichas estimaciones. De este modo, los ajustes del modelo se obtienen como la suma de estos efectos.

Los datos que hemos considerado procedían del modelo aditivo (1.5), por lo que los verdaderos efectos de cada una de las variables explicativas son conocidos y vienen dados por las funciones m_1, m_2, m_3 y m_4. Representamos estas funciones en la Figura 3.9, enfrentadas con sus respectivas estimaciones proporcionadas por el modelo. Tal y como cabría esperar, las estimaciones de los efectos se aproximan muy considerablemente a los verdaderos efectos de cada una de las variables explicativas.

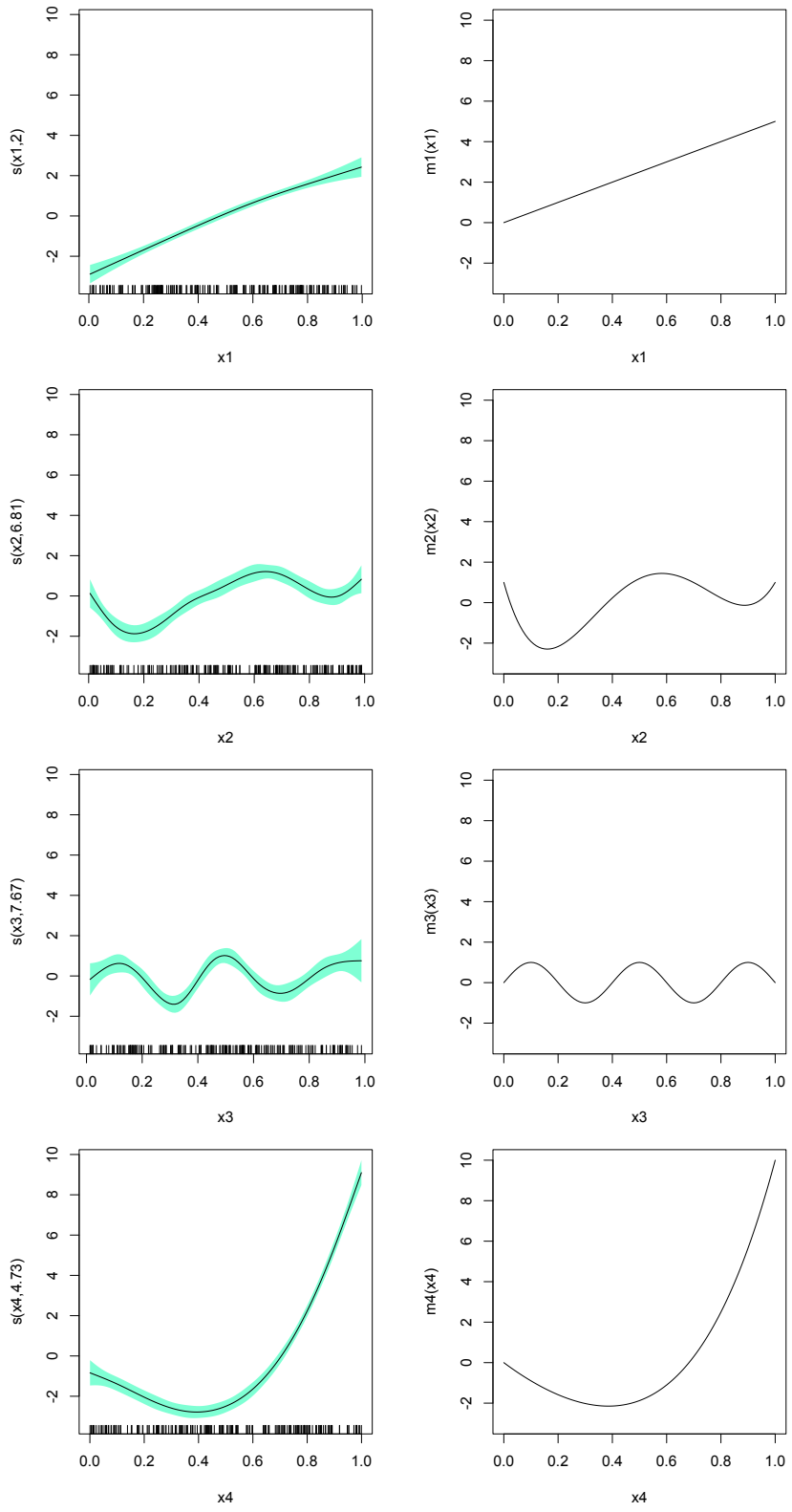
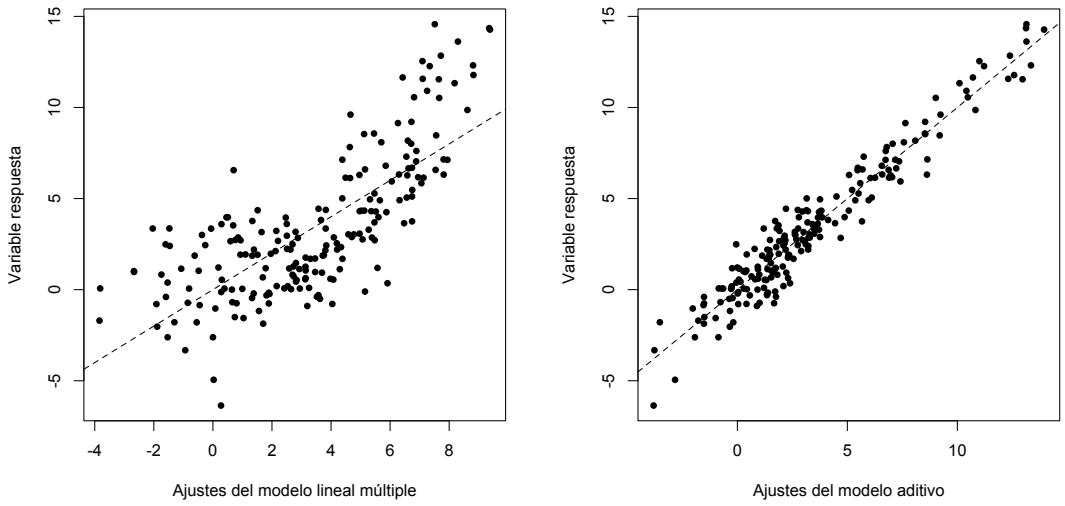


Figura 3.9: A la izquierda, estimaciones de los efectos de cada una de las cuatro variables explicativas sobre la variable respuesta asociadas al modelo `ma`. A la derecha, las representaciones de los verdaderos efectos poblacionales de cada una de ellas en el modelo (1.5).

Por último, vamos a realizar una comparación de la bondad del ajuste proporcionado por el modelo aditivo que acabamos de ajustar y del modelo lineal múltiple que planteamos en el Ejemplo 1.1, para los datos simulados a partir del modelo aditivo (1.5). Para ello consideramos una representación del valor de la respuesta frente a los ajustes de los dos respectivos modelos, como podemos ver en la Figura 3.10. La proximidad de los puntos a la diagonal, representada mediante una linea discontinua, es un indicador de la bondad del ajuste de un modelo. Comparando ambos gráficos no cabe duda de que los puntos están mucho más concentrados en torno a la diagonal en el caso del modelo aditivo, mientras que en el primer gráfico los puntos están mucho más dispersos. En consecuencia, podemos concluir que la estimación obtenida mediante el modelo aditivo es mucho mejor que la dada por el modelo lineal múltiple para nuestros datos, como era de esperar teniendo en cuenta la naturaleza de los datos simulados



(a) Modelo lineal múltiple.

(b) Modelo aditivo.

Figura 3.10: Representación de los valores reales frente a los ajustados por el modelo lineal múltiple (parte (a)) y por el modelo aditivo (parte (b)).

Capítulo 4

Aplicación a datos reales

A lo largo de este capítulo presentaremos un ejemplo de aplicación de los modelos aditivos en el análisis de datos reales. Los datos que usaremos se corresponden con mediciones diarias de la calidad del aire en Nueva York, realizadas de mayo a septiembre de 1973. Tenemos disponibles estos datos en el conjunto de datos `airquality` de  . En la base de datos se recogen mediciones diarias de las siguientes magnitudes:

1. `Ozone`: concentración de ozono media en ppb.

2. `Solar.R`: radiación solar en Langeys en la banda de frecuencia de 4000 a 7700 Angstroms.

3. `Wind`: velocidad media del viento en millas por hora.

4. `Temp`: temperatura máxima en grados Farenheit.

Los datos del ozono provienen del *Departamento de Conservación de Nueva York* y los datos meteorológicos del *Servicio Meterorológico Nacional*.

Vamos a considerar la concentración media de ozono como variable respuesta, Y, y tomaremos como variables explicativas la radiación solar, X_1, la velocidad media del viento, X_2, y la temperatura máxima, X_3.

A lo largo de este capítulo plantearemos el ajuste de un modelo de regresión lineal múltiple y de un modelo aditivo para los 104 primeros datos de `airquality` (una vez excluidos los datos faltantes). Posteriormente, mediremos la bondad de ambos ajustes viendo cuánto se aproximan las estimaciones de estos al valor de la variable respuesta para los restantes 7 datos del conjunto. Es decir, dividiremos la base de datos en una muestra de entrenamiento (para ajustar los modelos) y en una muestra de evaluación (para comparar ambos ajustes).

4.1. Ajuste de un modelo de regresión lineal múltiple

En primer lugar, podríamos pensar en ajustar un modelo de regresión lineal múltiple para representar la dependencia de la concentración del ozono en función de las variables explicativas, como hemos visto en el Capítulo 1, de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon.$$

Sin embargo, si representamos la concentración de ozono sobre cada una de las variables explicativas, como vemos en la Figura 4.1, podemos observar que estas no tienen un efecto lineal sobre la variable respuesta. Esto parece mostrar que la hipótesis de linealidad del modelo no se cumple.

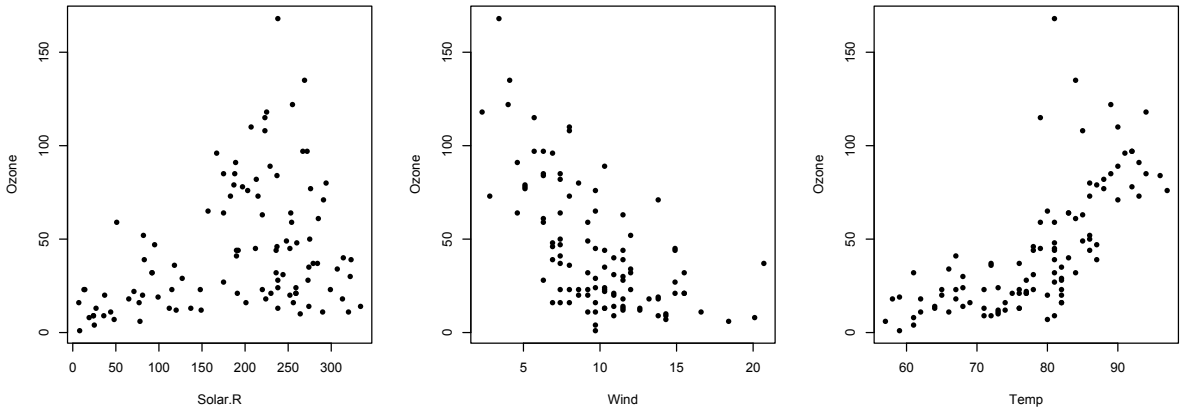


Figura 4.1: Representación de la variable respuesta sobre cada una de las variables explicativas asociadas a la base de datos `airquality`.

Aún así, si ajustamos un modelo lineal múltiple usando la función `lm` de se tendría que:

$$\widehat{Y} = -60.696 + 0.061 X_1 - 3.520 X_2 + 1.629 X_3 \tag{4.1}$$

Además, en el primer gráfico de la Figura 4.2, que representa los residuos frente a los ajustes del modelo lineal múltiple, podemos ver una cierta evolución que contradice la hipótesis de homocedasticidad. El segundo gráfico es un QQ-plot que claramente refleja un incumplimiento de la hipótesis de normalidad. Así, comprobamos que no se verifican las hipótesis básicas del modelo lineal múltiple. Consecuentemente, deducimos que este modelo no es correcto para representar esta situación.

Una vez hemos comprobado que no se cumplen las hipótesis de linealidad, normalidad y homocedastecidad del modelo lineal múltiple podemos formular un modelo alternativo o intentar

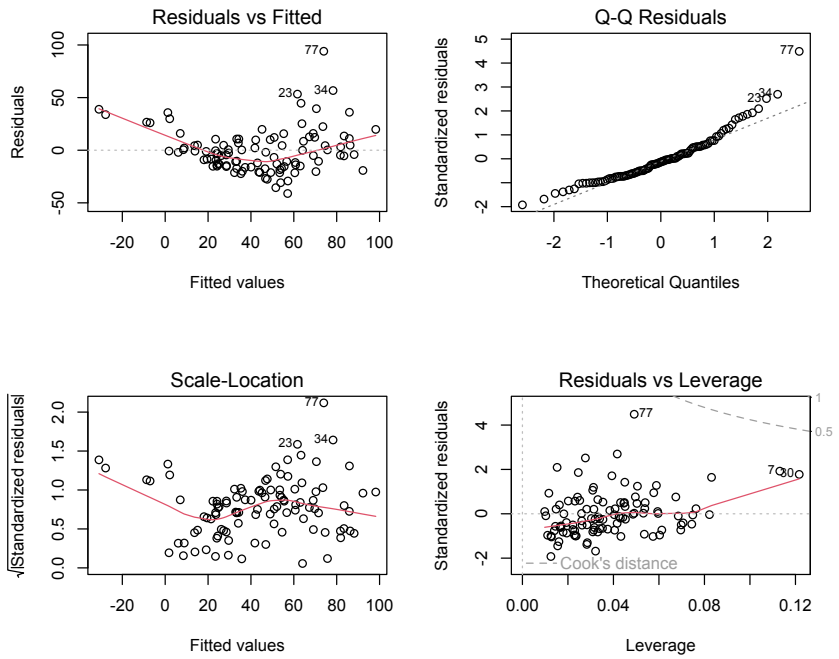


Figura 4.2: Gráficos para la validación del modelo de regresión lineal múltiple dado en (4.1) asociado a la base de datos `airquality`.

efectuar alguna transformación a los datos de modo que las hipótesis del modelo lineal múltiple se cumplan.

Puesto que nuestra variable respuesta `Ozone` es positiva, podemos aplicarle una transformación de Box-Cox, propuesta por Box y Cox (1964), que se define de la manera siguiente:

$$t_\alpha(Y) = \begin{cases} \dfrac{Y^\alpha - 1}{\alpha}, & \text{si } \alpha \neq 0, \\ \ln(Y), & \text{si } \alpha = 0. \end{cases}$$

El parámetro α óptimo de la transformación de Box-Cox es aquel para el que obtengamos la mejor aproximación a una distribución normal para nuestra variable respuesta. Seleccionaremos el parámetro α que maximice la función de verosimilitud, que en nuestro caso es $\alpha = 0.222$.

Tomando este α óptimo, efectuamos una transformación Box-Cox a los datos, y ajustamos de nuevo un modelo lineal múltiple para los datos transformados, que tendrá la forma:

$$t_\alpha(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon. \tag{4.2}$$

Siguiendo el mismo procedimiento que antes, representamos la variable respuesta en función

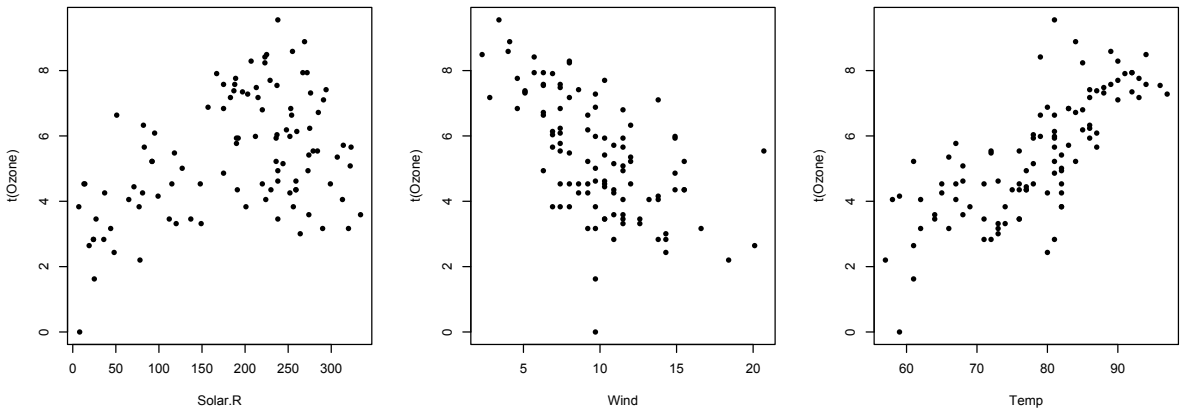


Figura 4.3: Representación de la variable respuesta transformada mediante la transformación Box-Cox sobre cada una de las variables explicativas asociadas a la base de datos `airquality`.

de cada una de las variables explicativas en la Figura 4.3, de modo que observamos que los efectos de cada una de las variables explicativas sobre la concentración de ozono siguen sin ser lineales.

Aún así el ajuste sobre el modelo transformado resultaría:

$$\widehat{t_\alpha(Y)} = -2.002 + 0.004X_1 - 0.154X_2 + 0.102X_3 \tag{4.3}$$

Observando la Figura 4.4, que recoge la validación del modelo (4.3), podemos ver que hemos solucionado los problemas con la hipótesis de normalidad que teníamos para el modelo (4.1), pues los puntos del QQ-plot se aproximan considerablemente a una linea recta. Sin embargo, el gráfico de residuos frente a ajustes parece seguir reflejando una cierta evolución, de forma que tampoco podemos afirmar que se cumplen las hipótesis de linealidad y homocedasticidad para el modelo transformado (4.3). En consecuencia, debemos pensar en ajustar un modelo alternativo.

4.2. Ajuste de un modelo de regresión aditivo

Seguimos considerando los datos transformados mediante la transformación Box-Cox de parámetro $\alpha = 0.222$, y ajustamos ahora un modelo aditivo a partir de ellos, como vimos en el Capítulo 3. Tendremos así un modelo de la forma:

$$t_\alpha(Y) = \beta_0 + m_1(X_1) + m_2(X_2) + m_3(X_3) + \varepsilon. \tag{4.4}$$

Por lo tanto, ajustamos el siguiente modelo aditivo usando el software estadístico R:

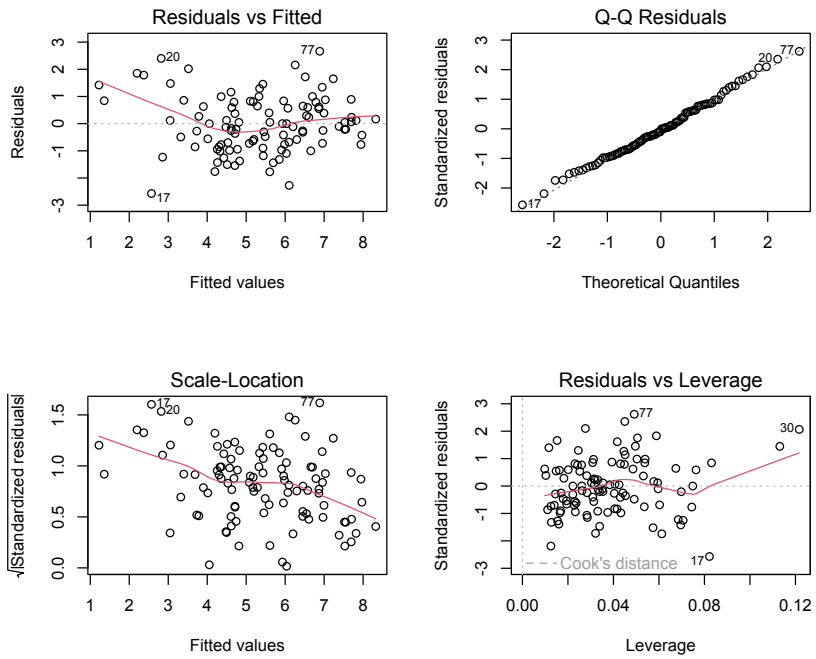


Figura 4.4: Gráficos para la validación del modelo de regresión lineal múltiple (4.3) para los datos transformados mediante una transformación Box-Cox asociados a la base de datos `airquality`.

```
> ma1 <- gam(Ozone.t ~ s(Solar.R) + s(Wind) + s(Temp))
```

De este modo, los efectos estimados de la radiación solar, el viento y la temperatura sobre la concentración de ozono serían respectivamente los representados en la Figura 4.5, y los ajustes dados por el modelo `ma1` serán la suma de estos efectos.

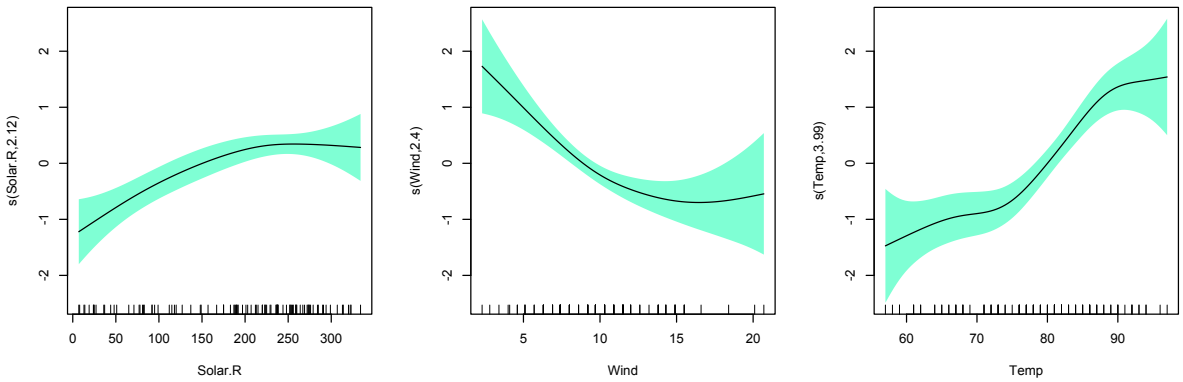


Figura 4.5: Efectos de las variables explicativas estimados por el modelo aditivo `ma1` asociado a la base de datos `airquality`.

Una vez comprobamos que el modelo lineal múltiple no era correcto para representar estos datos, podemos ver que en este caso el modelo aditivo es una buena alternativa. Su principal ventaja es la interpretabilidad de los efectos de las covariables.

Observando la Figura 4.5 resulta sencillo interpretar los efectos de cada una de las variables explicativas sobre la variable respuesta. De este modo, podemos ver que la concentración de ozono aumenta ligeramente a medida que aumenta la radiación solar. Además, según aumenta la temperatura máxima el aumento en la concentración del ozono es todavía mayor. Por el contrario, a medida que se incrementa la velocidad del viento la concentración de ozono disminuye.

Una vez hemos ajustado el modelo aditivo procedemos a realizar la validación del mismo. Para ello empleamos la función `gam.check`, que nos devuelve lo siguiente:

```
> gam.check(ma1)

Method: GCV   Optimizer: magic
Smoothing parameter selection converged after 6 iterations.
The RMS GCV score gradient at convergence was 5.425529e-06 .
The Hessian was positive definite.
Model rank =  28 / 28

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

             k'  edf k-index p-value
s(Solar.R) 9.00 2.52    1.05    0.61
s(Wind)    9.00 2.55    0.93    0.26
s(Temp)    9.00 4.29    0.97    0.38
```

Además de esto, el comando `gam.check` también nos proporciona una serie de gráficos básicos para la validación del modelo, que podemos ver representados en la Figura 4.6.

Observando estos gráficos podemos concluir que se verifican las hipótesis del modelo: el QQ-plot refleja claramente la normalidad de los datos y el gráfico de residuos sobre ajustes parece mostrar que la varianza se mantiene constante con la media. Además, observando el gráfico inferior derecho de valores reales frente a ajustados notamos que los puntos se distribuyen uniformemente en torno a la diagonal y se concentran en torno a esta considerablemente, sugiriendo así que se trata de un modelo adecuado para representar estos datos.

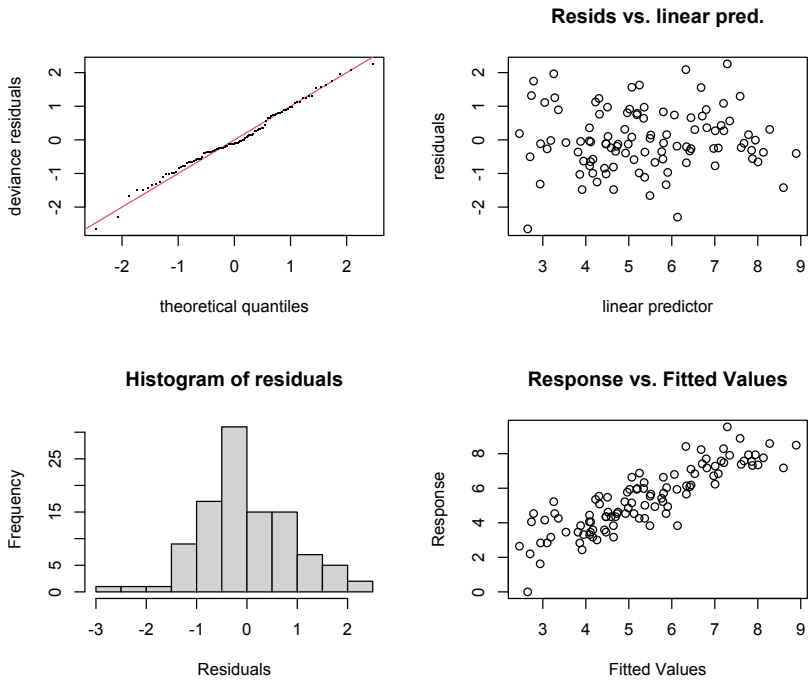


Figura 4.6: Gráficos básicos para la validación del modelo aditivo `ma1` obtenidos mediante la función `gam.check` para la base de datos `airquality`.

4.3. Comparación de ambos ajustes

Como hemos mencionado al principio de este capítulo, para el ajuste de estos modelos hemos empleado únicamente los 104 primeros datos disponibles dentro de la base de datos `airquality`. Con la finalidad de comparar la bondad del ajuste del modelo aditivo (4.4) y del modelo lineal múltiple (4.2), calcularemos las predicciones de ambos modelos para los 7 datos restantes.

Observamos que ambos modelos nos proporcionarán estimaciones correspondientes a la variable respuesta transformada $t_\alpha(\texttt{Ozone})$. De esta forma, una vez obtenidas estas predicciones, desharemos la transformación Box-Cox obteniendo así estimaciones para la variable respuesta original `Ozone`. Así, podemos medir la proximidad de estas estimaciones a los valores reales de la variable respuesta. Para ello emplearemos como criterio de error el error cuadrático medio de predicción, que vendrá dado por:

$$\frac{1}{7}\sum_{i=1}^{7}(Y_i - \widehat{m}(X_i))^2,$$

y el error absoluto medio de predicción, dado por:

$$\frac{1}{7}\sum_{i=1}^{7}|Y_i - \widehat{m}(X_i)|.$$

En la Tabla 4.1 mostramos el error cuadrático medio de predicción y el error absoluto medio de predicción asociados al modelo lineal múltiple (4.2) y al modelo aditivo (4.4) ajustados sobre los datos `airquality`. Se observa que tanto el error cuadrático medio de predicción como el error cuadrático absoluto de predicción obtenidos son menores para el caso del modelo aditivo (4.4) que para el del lineal múltiple (4.2).

	Modelo lineal múltiple	Modelo aditivo
Error cuadrático medio de predicción	52.515	32.259
Error cuadrático absoluto de predicción	5.740	4.916

Tabla 4.1: Errores de predicción del modelo lineal múltiple (4.2) y del modelo aditivo (4.4) para la muestra de evaluación considerada relativa a la base de datos `airquality`.

De esta forma, concluimos que la estimación dada por el modelo aditivo es mejor que la que obtendríamos en el caso de emplear un modelo lineal múltiple. Así, acabamos de ver que para la representación de estos datos resultaría adecuado usar un modelo aditivo, proponiendo así un ejemplo de aplicación de este tipo de modelos a un conjunto de datos reales.

Capítulo 5

Conclusiones

En el Capítulo 1 de este documento realizamos una introducción a los modelos de regresión como forma de representar la relación de dependencia de una variable respuesta Y respecto a las variables explicativas $X_1, \ldots, X_p$. Además, planteamos como primera aproximación el modelo lineal múltiple, un modelo de regresión paramétrica, cuya formulación recordamos que venía dada por:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \tag{5.1}$$

donde $\varepsilon \in N(0, \sigma^2)$.

Cuando planteamos un modelo de este tipo, podemos diferenciar cuál es el efecto de cada una de las variables explicativas sobre la variable respuesta e interpretarlo: si aumentamos el valor de X_j en una unidad, y si el resto de covariables se mantienen constantes, la variable respuesta aumentará en β_j unidades. Sin embargo, observamos que estos efectos se están suponiendo lineales, lo cual no siempre se cumple en la práctica. Es por esto que en este trabajo planteamos una alternativa a este tipo de modelos: los modelo aditivos.

Como decíamos, en la práctica no siempre podemos asegurar el cumplimiento de las hipótesis básicas del modelo lineal múltiple. Ilustramos este hecho en el Capítulo 4 para el conjunto de datos reales `airquality`, ya que mostramos que el modelo lineal múltiple resultaría incorrecto para representar esta situación.

En este tipo de casos es necesario recurrir a métodos no paramétricos. Es decir, consideraríamos modelos de la forma:

$$Y = m(X_1, \ldots, X_p) + \varepsilon, \tag{5.2}$$

donde la función m es totalmente desconocida. Un modelo no paramétrico nos da una mayor flexibilidad, ya que no asumimos ningún tipo de restricción sobre la forma del modelo de regresión.

En el Capítulo 2 estudiamos varias formas de ajustar este tipo de modelos para el caso

univariante. Así, planteamos un modelo no paramétrico de la forma

$$Y = m(X) + \varepsilon,$$

y estudiamos su ajuste mediante estimadores tipo núcleo: constante local y lineal local. En este tipo de situaciones será fundamental seleccionar correctamente el parámetro de suavizado puesto que su efecto será determinante en la bondad del ajuste.

Volviendo al caso multivariante, si estimásemos $m(X_1, \ldots, X_p)$ para el modelo (5.2) desconoceríamos en qué manera influye cada una de las variables explicativas en la variable respuesta, suponiendo una desventaja respecto a los modelos paramétricos que sí nos permiten una interpretabilidad de los efectos de cada covariable. Además, en este contexto multivariante la elección de los parámetros de suavizado también se complica notablemente.

Así, nos interesa considerar un modelo no paramétrico, pero que simplifique la expresión general dada en (5.2). En este contexto presentamos los modelos aditivos, que nos permiten escribir la variable resuesta de la siguiente forma:

$$Y = \beta_0 + \sum_{j=1}^{p} m_j(X_j) + \varepsilon. \tag{5.3}$$

A lo largo del Capítulo 3 estudiamos la estimación de este tipo de modelos, empleando bases de *splines* para representar cada uno de los efectos m_j del modelo (5.3) y poder estimar el modelo por mínimos cuadrados penalizados, seleccionando los parámetros de suavizado mediante criterios de validación cruzada.

Observamos que estamos tomando como hipótesis la aditividad de los efectos de las variables explicativas. Entonces, no todas las situaciones se pueden representar con un modelo aditivo, pero aún así, el grado de flexibilidad que nos proporcionan sigue siendo muy elevado. Por otra parte, es la hipótesis de aditividad la que nos permite interpretar los efectos de cada una de las covariables sobre el ajuste de manera independiente.

Retomamos de nuevo el análisis realizado en el Capítulo 4 de la base de datos reales `airquality`, para los que ajustamos un modelo aditivo de la forma (5.3) con el objetivo de explicar el comportamiento de la variable `Ozone`. Las estimaciones dadas por este modelo para los efectos m_1, m_2 y m_3 asociados a las variables `Solar.R`, `Wind` y `Temp` respectivamente son las representadas en la Figura 5.1.

Como ya hemos dicho, el modelo lineal múltiple resultaba inadecuado para representar estos datos, por no cumplirse las hipótesis básicas del modelo. Sin embargo, el modelo aditivo planteado, cuyas hipótesis son mucho menos restrictivas, logra representar bien esta situación. Además,

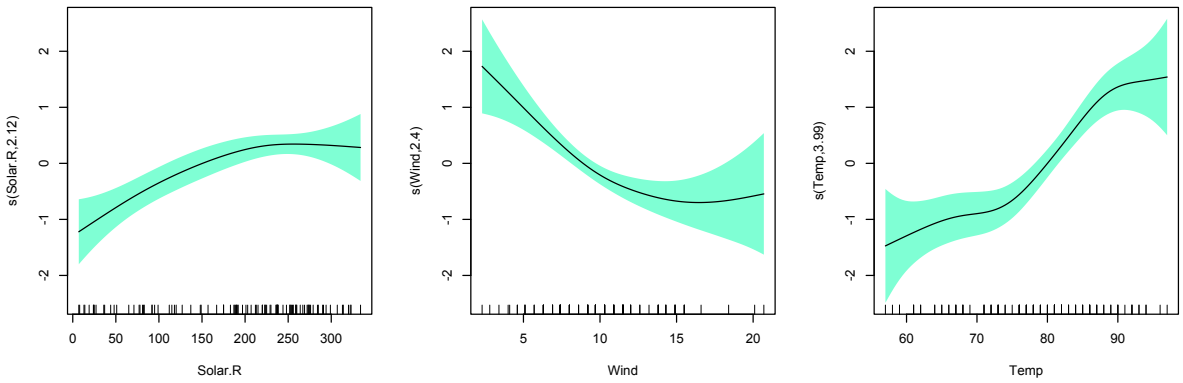


Figura 5.1: Efectos de las variables explicativas estimados por el modelo aditivo `ma1` asociado a la base de datos `airquality`.

nos permite interpretar el efecto que tiene cada una de las variables explicativas sobre la variable respuesta observando la Figura 5.1, tal y como hemos visto en el Capítulo 4.

En conclusión, los modelos aditivos nos dan la flexibilidad de un modelo no paramétrico pero manteniendo la interpretabilidad de los efectos de las variables explicativas característica del modelo lineal múltiple. Es por esto que presentamos los modelos aditivos como un buen "punto intermedio" entre el modelo lineal múltiple (5.1) y el modelo no paramétrico (5.2).

Anexo I

Código de R desarrollado

I.1. Datos simulados a partir del modelo aditivo (1.5)

En esta sección recogemos el código de  que empleamos para la simulación de datos a partir del modelo aditivo planteado en (1.5) y su posterior análisis. Comenzamos ajustando un modelo lineal múltiple para la estimación de estos datos en el Capítulo 1, pero como vimos, el modelo resultó incorrecto. En la Sección 3.3 retomamos estos datos para ilustrar el ajuste de un modelo aditivo, y compararlo con el modelo lineal múltiple anterior.

```
#--- Fijamos semilla
set.seed(12345)

#--- Efectos de las variables explicativas
m1 <- function(x){return(5*x)}
theta <- c(1,-48,218,-315,145)
m2 <- function(x){return(theta[1]+theta[2]*x+theta[3]*x^2+theta[4]*x^3
+theta[5]*x^4)}
m3 <- function(x){return(sin(5*pi*x))}
m4 <- function(x){return(10*(x^4+x^2-x))}

#--- Representación de los efectos
par(mfrow=c(2,2))
curve(m1, from = 0, to = 1, ylim=c(-3,10), xlab="x1", ylab="m1(x1)")
curve(m2, from = 0, to = 1, ylim=c(-3,10), xlab="x2", ylab="m2(x2)")
curve(m3, from = 0, to = 1, ylim=c(-3,10), xlab="x3", ylab="m3(x3)")
curve(m4, from = 0, to = 1, ylim=c(-3,10), xlab="x4", ylab="m4(x4)")
```

```
#--- Simulación de los datos
n=200
error=rnorm(n)

x1=runif(n)
f1=5*x1

x2=runif(n)
f2=theta[1]+theta[2]*x2+theta[3]*x2^2+theta[4]*x2^3+theta[5]*x2^4

x3=runif(n)
f3=sin(5*pi*x3)

x4=runif(n)
f4=10*(x4^4+x4^2-x4)

y=f1+f2+f3+f4+error


## MODELO LINEAL MÚLTIPLE ##

#--- Estimación del modelo lineal múltiple
mlm=lm(y~x1+x2+x3+x4)
summary(mlm)

#--- Gráficos para la validación del modelo
par(mfrow=c(2,2))
plot(mlm)

#--- Valores de la variable respuesta frente a los ajustados por el modelo
plot(y~mlm$fit, pch=16, cex=0.9, xlab="Ajustes del modelo",
ylab="Variable respuesta")
abline(0,1,lty=2) # diagonal


## MODELO ADITIVO ##
```

```
library(mgcv)

#--- Ajuste de un modelo aditivo con una base de k=10 splines cúbicos
ma <- gam(y ~ s(x1,k=10,bs="cr") + s(x2,k=10,bs="cr") + s(x3,k=10,bs="cr")
+ s(x4,k=10,bs="cr"))
summary(ma)

#--- Ajuste de un modelo aditivo por defecto
ma0 <- gam(y ~ s(x1) + s(x2) + s(x3) + s(x4))
summary(ma0)

#--- Efectos de cada variable explicativa estimados por el modelo
plot(ma, shade=TRUE, shade.col = "aquamarine")
plot(ma, shade=TRUE, shade.col = "aquamarine")
plot(ma, shade=TRUE, shade.col = "aquamarine")
plot(ma, shade=TRUE, shade.col = "aquamarine")

#--- Verdaderos efectos de las variables explicativas
curve(m1, from = 0, to = 1, ylim=c(-3,10), xlab="x1", ylab="m1(x1)")
curve(m2, from = 0, to = 1, ylim=c(-3,10), xlab="x2", ylab="m2(x2)")
curve(m3, from = 0, to = 1, ylim=c(-3,10), xlab="x3", ylab="m3(x3)")
curve(m4, from = 0, to = 1, ylim=c(-3,10), xlab="x4", ylab="m4(x4)")

#--- Variable respuesta frente a los ajustes del modelo aditivo
plot(y~mlm$fit, pch=16, cex=0.9,xlab="Ajustes del modelo lineal múltiple",
ylab="Variable respuesta")
abline(0,1,lty=2) # diagonal

#--- Variable respuesta frente a los ajustes del modelo lineal múltiple
plot(y~ma$fit, pch=16, cex=0.9, xlab="Ajustes del modelo aditivo",
ylab="Variable respuesta")
abline(0,1,lty=2) # diagonal
```

I.2. Datos simulados a partir del modelo univariante (2.1)

Consideramos ahora el código de R correspondiente a la simulación y análisis de los datos generados a partir del modelo univariante (2.1), dado por:

$$Y = m(X) + \varepsilon = 1 - 48X + 218X^2 - 315X^3 + 145X^4 + \varepsilon.$$

En el Capítulo 2 ajustamos varios modelos de regresión lineal local tomando distintos parámetros de suavizado para estos datos, reflejando la relevancia de este parámetro en el ajuste. Además, en el Capítulo 3 empleamos estos datos para ilustrar la estimación basada en una base lineal a trozos y distintos *splines*.

```
#--- Simulación de los datos
error2=rnorm(n)
y2=f2+error2


## MODELO LINEAL LOCAL ##

library(KernSmooth)

#--- Estimación infrasuavizada
#--- Ajuste lineal local para h=0.005
infrasuavizado <- locpoly(x2, y2, bandwidth = 0.005)
plot(y2~x2, pch=16, cex=0.9, ylab="Y", xlab="X")
curve(m2, from = 0, to = 1, col = 1, lwd=1.2, add = TRUE)
lines(infrasuavizado,col="red", lwd=1.2)

#--- Estimación sobresuavizada
#--- Ajuste lineal local para h=0.5
sobresuavizado <- locpoly(x2, y2, bandwidth = 0.5)
plot(y2~x2, pch=16, cex=0.9, ylab="Y",xlab="X")
curve(m2, from = 0, to = 1, col = 1, lwd=1.2, add = TRUE)
lines(sobresuavizado,col="blue",lwd=1.2)

#--- Obtención del parámetro ventana óptimo
h <- dpill(x2,y2)

#--- Ajuste lineal local para h óptimo
```

```
optimo <- locpoly(x2, y2, bandwidth = h)
plot(y2~x2, pch=16, cex=0.9, ylab="Y",xlab="X")
curve(m2, from = 0, to = 1, col = 1, lwd=1.2, add = TRUE)
lines(optimo,col="green3", lwd=1.2)


## AJUSTE LINEAL A TROZOS ##

#--- Construcción de la j-ésima función de la base de funciones lineales a
#--- trozos definida por los nodos z
b <- function(x,z,j) {
dj <- z*0
dj[j] <- 1
approx(z,dj,x)$y  # interpolación lineal
}

#--- Construcción de la matriz de diseño, donde Xij=bj(xi), para los datos x
#--- y los nodos z
b.X <- function(x,z) {
k <- length(z)
n <- length(x)
X <- matrix(NA,n,k)
for (j in 1:k) X[,j] <- b(x,z,j)
X
}

#--- Representación de los datos
plot(y2~x2, pch=16, cex=0.9, ylab="Y",xlab="X")

#--- Verdadera función de regresión
curve(m2, from = 0, to = 1, col = 1, lwd=1.2, add = TRUE)

#--- Ajuste lineal a trozos para k=3
#--- Nodos
z1 <- seq(min(x2), max(x2), length=3)

#--- Estimación del modelo por mínimos cuadrados
X <- b.X(x2,z1)
```

```
modelo <- lm(y2 ~ X - 1)
beta <- coef(modelo)

#--- Matriz de predicción
pred <- seq(min(x2), max(x2), length=200)
X.pred <- b.X(pred,z1)

#--- Representación de la estimación del modelo
lines(pred,X.pred %*% beta, col="blue", lwd=1.2)

#--- Ajuste lineal a trozos para k=10
z2 <- seq(min(x2),max(x2),length=10)
X <- b.X(x2,z2)
modelo <- lm(y2 ~ X - 1)
beta <- coef(modelo)
X.pred <- b.X(pred,z2)
lines(pred,X.pred %*% beta, col="green3", lwd=1.2)

#--- Ajuste lineal a trozos para k=40
z3 <- seq(min(x2),max(x2),length=40)
X <- b.X(x2,z3)
modelo <- lm(y2 ~ X - 1)
beta <- coef(modelo)
X.pred <- b.X(pred,z3)
lines(pred,X.pred %*% beta, col="red", lwd=1.2)

legend("bottomright", legend=c("Función de regresión teórica","k=3","k=10",
"k=40"), col=c(1,"blue","green3","red"), lty=1, cex=0.8)


## AJUSTE LINEAL A TROZOS PENALIZADO ##

#--- Ajuste de un modelo lineal a trozos penalizado
ajuste.ltp <- function(y,x,z,lambda) {
X <- b.X(x,z) # matriz de diseño
D <- diff(diag(length(z)),differences=2) # raíz de la matriz de penalización
X.aum <- rbind(X,sqrt(lambda)*D) # matriz de diseño aumentada
y.aum <- c(y,rep(0,nrow(D))) # vector de respuestas aumentado
```

```
lm(y.aum ~ X.aum - 1) # estimación del modelo por mínimos cuadrados
}

#--- Nodos
z <- seq(min(x2), max(x2), length=20)

#--- Ajuste de un modelo lineal a trozos penalizado con lambda=0.1
#--- Parámetro de suavizado
lambda <- 0.1

#--- Estimación del modelo
modelo <- ajuste.ltp(y2,x2,z,lambda)
beta <- coef(modelo)

#--- Representación de los datos
plot(x2, y2, pch=16, cex=0.9, ylab="Y", xlab="X")

#--- Verdadera función de regresión
curve(m2, from = 0, to = 1, col = 1, lwd=1.2, add = TRUE)

#--- Matriz de predicción
pred <- seq(min(x2), max(x2), length=200)
X.pred <- b.X(pred,z)

#--- Representación de la estimación
lines(pred,X.pred %*% beta,col="red")

#--- Ajuste de un modelo lineal a trozos penalizado con lambda=1000
lambda <- 1000
modelo <- ajuste.ltp(y2,x2,z,lambda)
beta <- coef(modelo)
plot(x2, y2, pch=16, cex=0.9, ylab="Y", xlab="X")
curve(m2, from = 0, to = 1, col = 1, lwd=1.2, add = TRUE)
X.pred <- b.X(pred,z)
lines(pred,X.pred %*% beta,col="blue")


## SELECCIÓN DEL PARÁMETRO DE SUAVIZADO MEDIANTE GCV ##
```

```
#--- Rejilla de valores del parámetro de suavizado
rho <- seq(-10,20,length=100)

n <- length(y2)
GCV <- rep(NA,100)
z <- seq(min(x2),max(x2),length=50)
X.pred <- b.X(pred,z)

#--- Valor de la GCV para cada parámetro de suavizado
for (i in 1:100) {
modelo <- ajuste.ltp(y2,x2,z,exp(rho[i])) # ajuste del modelo por mínimos
cuadrados penalizados
tr.H <- sum(influence(modelo)$hat[1:n]) # traza de la matriz hat del modelo
rss <- sum((y2-fitted(modelo)[1:n])^2) # RSS
GCV[i] <- n*rss/(n-tr.H)^2 # valor de la puntuación GCV
}

#--- Valor de la GCV frente al logaritmo de lambda
plot(rho, GCV, type="l", xlab=expression(log(lambda)),
ylab=expression(GCV(lambda)))

#--- Parámetro de suavizado óptimo
lambda <- exp(rho[GCV==min(GCV)]); lambda

#--- Ajuste de un modelo lineal a trozos penalizado con lambda óptimo
modelo <- ajuste.ltp(y2,x2,z,lambda)
beta <- coef(modelo)

#--- Representación
plot(x2, y2, pch=16, cex=0.9, ylab="Y",xlab="X")
curve(m2, from = 0, to = 1, col = 1, lwd=1.2, add = TRUE)
lines(pred,X.pred %*% beta,col="green3",lwd=1.2)


## AJUSTE MEDIANTE DISTINTOS SPLINES DE SUAVIZADO ##

library(mgcv)
```

```
datos=data.frame(y,x1,x2,x3,x4)

plot(y2~x2, pch=16, cex=0.9, ylab="Y",xlab="X")
curve(m2, from = 0, to = 1, col = 1, lwd=1.2, add = TRUE)

#--- Spline cúbico
sm <- smoothCon(s(x2, bs="cr"), data=datos, knots=NULL)[[1]]
beta <- coef(lm(y2 ~ sm$X-1))
x2.pred <- seq(0, 1, length=200)
Xp <- PredictMat(sm, data.frame(x2=x2.pred))
lines(x2.pred, Xp%*%beta, col="red", lwd=1.2)

#--- TPS
sm <- smoothCon(s(x2, bs="tp"), data=datos, knots=NULL)[[1]]
beta <- coef(lm(y2 ~ sm$X-1))
x2.pred <- seq(0, 1, length=200)
Xp <- PredictMat(sm, data.frame(x2=x2.pred))
lines(x2.pred, Xp%*%beta, col="blue", lwd=1.2)

#--- Spline cúbico cíclico
sm <- smoothCon(s(x2, bs="cc"), data=datos, knots=NULL)[[1]]
beta <- coef(lm(y2 ~ sm$X-1))
x2.pred <- seq(0, 1, length=200)
Xp <- PredictMat(sm, data.frame(x2=x2.pred))
lines(x2.pred, Xp%*%beta, col="green3", lwd=1.2)

#--- P-spline
sm <- smoothCon(s(x2, bs="ps"), data=datos, knots=NULL)[[1]]
beta <- coef(lm(y2 ~ sm$X-1))
x2.pred <- seq(0, 1, length=200)
Xp <- PredictMat(sm, data.frame(x2=x2.pred))
lines(x2.pred, Xp%*%beta, col="hotpink", lwd=1.2)

legend("bottomright", legend=c("Función de regresión teórica","Spline cúbico",
"TPS", "Spline cúbico cíclico", "P-spline"), col=c(1,2,3,4,6), lty=1, cex=0.75,
lwd=1.2)
```

I.3. Datos de `airquality`

Incluimos a continuación el código de  empleado en el Capítulo 4 para el ajuste de los modelos lineal múltiple y aditivo a partir de los datos de `airquality`, y la posterior comparación de ambos.

```
#--- Carga de datos
data("airquality")
datos0 <- na.omit(airquality)[,c("Ozone", "Solar.R", "Wind", "Temp")]
rownames(datos0) <- 1:nrow(datos0)

#--- Consideramos los 104 primeros datos para el ajuste
datos1 <- datos0[1:104,]
datos2 <- datos0[105:111,]

Ozone <- datos1$Ozone
Solar.R <- datos1$Solar.R
Wind <- datos1$Wind
Temp <- datos1$Temp


## MODELO LINEAL MÚLTIPLE ##

#--- Ajustamos un modelo lineal múltiple
mlm <- lm(Ozone ~ Solar.R + Wind + Temp)
summary(mlm)

#--- Gráficos para la validación
par(mfrow=c(2,2))
plot(mlm)

#--- Variable respuesta sobre cada una de las explicativas
par(mfrow=c(1,3))
plot(Ozone~Solar.R, pch=16, cex=0.9)
plot(Ozone~Wind, pch=16, cex=0.9)
plot(Ozone~Temp, pch=16, cex=0.9)
```

```
## MODELO LINEAL MÚLTIPLE TRANSFORMADO ##

library(MASS)

#--- Parámetro óptimo para la transformación Box-Cox
transformacion.boxcox <- boxcox(Ozone ~ Solar.R + Wind + Temp)
alpha <- transformacion.boxcox$x[which.max(transformacion.boxcox$y)]

#--- Datos transformados
Ozone.t <- (Ozone^{alpha}-1)/alpha

#--- Ajuste de un modelo lineal múltiple para los datos transformados
mlm2 <- lm(Ozone.t ~ Solar.R + Wind + Temp)
summary(mlm2)

#--- Gráficos para la validación
par(mfrow=c(2,2))
plot(mlm2)

#--- Variable respuesta sobre cada una de las explicativas
par(mfrow=c(1,3))
plot(Ozone.t~Solar.R,pch=16,cex=0.9,ylab="t(Ozone)")
plot(Ozone.t~Wind,pch=16,cex=0.9,ylab="t(Ozone)")
plot(Ozone.t~Temp,pch=16,cex=0.9,ylab="t(Ozone)")


## MODELO ADITIVO ##

library(mgcv)
ma1 <- gam(Ozone.t ~ s(Solar.R) + s(Wind) + s(Temp))
summary(ma1)

#--- Validación del modelo
gam.check(ma1)

#--- Efectos de cada variable explicativa estimados por el modelo
par(mfrow=c(1,3))
plot(ma1, shade=TRUE, shade.col = "aquamarine")
```

```
## PREDICCIONES ##

Solar.R.pred <- datos2$Solar.R
Wind.pred <- datos2$Wind
Temp.pred <- datos2$Temp

#--- Predicciones del modelo aditivo
ma1.pred <- predict(ma1, newdata=data.frame(Solar.R=Solar.R.pred, Wind=Wind.pred,
Temp=Temp.pred), type="response")

#--- Predicciones del modelo lineal múltiple
mlm2.pred <- predict(mlm2,newdata=data.frame(Solar.R=Solar.R.pred, Wind=Wind.pred,
Temp=Temp.pred))

#--- Valores reales de la variable respuesta
Ozone.real <- datos2$Ozone

#--- Deshacemos la transformación Box-Cox de las estimaciones
ma1.pred2 <- (ma1.pred*alpha+1)^(1/alpha)
mlm2.pred2 <- (mlm2.pred*alpha+1)^(1/alpha)

#--- MSE
mean((Ozone.real-mlm2.pred2)^2)
mean((Ozone.real-ma1.pred2)^2)

#--- MAE
mean(abs(Ozone.real-mlm2.pred2))
mean(abs(Ozone.real-ma1.pred2))
```

I.4. Otras representaciones

Finalmente, a lo largo de esta sección presentaremos diferentes representaciones gráficas que hemos presentado a lo largo del documento.

```
## REPRESENTACIÓN DE UNA BASE LINEAL A TROZOS CON K=6 ##
```

```
z <- seq(0, 1, len=6)

y1 <- c(1,0,0,0,0,0)
y2 <- c(0,1,0,0,0,0)
y3 <- c(0,0,1,0,0,0)
y4 <- c(0,0,0,1,0,0)
y5 <- c(0,0,0,0,1,0)
y6 <- c(0,0,0,0,0,1)

par(mfrow=c(2,3))
plot(z, y1, type="l", ylab="b1(x)")
abline(h=0, lty=2)
abline(v=z, lty=3, col="grey")
plot(z, y2, type="l", ylab="b2(x)")
abline(h=0, lty=2)
abline(v=z, lty=3, col="grey")
plot(z, y3, type="l", ylab="b3(x)")
abline(h=0, lty=2)
abline(v=z, lty=3, col="grey")
plot(z, y4, type="l", ylab="b4(x)")
abline(h=0, lty=2)
abline(v=z, lty=3, col="grey")
plot(z, y5, type="l", ylab="b5(x)")
abline(h=0, lty=2)
abline(v=z, lty=3, col="grey")
plot(z, y6, type="l", ylab="b6(x)")
abline(h=0, lty=2)
abline(v=z, lty=3, col="grey")


## REPRESENTACIÓN DE UN SPLINE CÚBICO NATURAL ##

#--- Puntos
x <- seq(0,1,len=6)
y <- c(0.1,0.4,0.2,0.9,0.4,0.6)

library(pracma)
```

```
#--- Spline cúbico natural interpolando estos puntos
xs <- seq(0,1,len=100)
ys <- cubicspline(x, y, xs, endp2nd = TRUE)

#--- Representación del spline
plot(x, y, ylim=c(0,1), pch=16, cex=0.9, ylab="s(x)")
lines(xs, ys)


## REPRESENTACIÓN DE UNA BASE DE SPLINES CÚBICOS NATURALES CON K=6 ##

library(splines2)
packageVersion("splines2")

z <- seq(0, 1, len=6)
x <- seq(0, 1, 0.01)
nskMat <- nsk(x, knots = c(0.2,0.4,0.6,0.8), intercept = TRUE)

plot(nskMat, ylab = "nsk()", mark_knots = "all")

par(mfrow=c(2,3))
plot(x, nskMat[,1], type="l", ylab="b1(x)")
abline(h=0, lty=2)
abline(v=z, lty=3, col="grey")
plot(x, nskMat[,2], type="l", ylab="b2(x)")
abline(h=0, lty=2)
abline(v=z, lty=3, col="grey")
plot(x, nskMat[,3], type="l", ylab="b3(x)")
abline(h=0, lty=2)
abline(v=z, lty=3, col="grey")
plot(x, nskMat[,4], type="l", ylab="b4(x)")
abline(h=0, lty=2)
abline(v=z, lty=3, col="grey")
plot(x, nskMat[,5], type="l", ylab="b5(x)")
abline(h=0, lty=2)
abline(v=z, lty=3, col="grey")
plot(x, nskMat[,6], type="l", ylab="b6(x)")
```

```
abline(h=0, lty=2)
abline(v=z, lty=3, col="grey")
```

Referencias

Box, G. E. P., y Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211-252.

Fan, J. (1992). Design-adaptive Nonparametric Regression. *Journal of the American Statistical Associattion*, *87*(420), 998-1004.

Fan, J., y Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall.

Faraway, J. J. (2005). *Linear Models wih R*. Chapman & Hall.

Green, P. J., y Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach*. Chapman & Hall.

Hastie, T. J. (2023). Generalized Additive Models. Paquete de R. Descargado de https://cran.r-project.org/web/packages/gam/gam.pdf

Hastie, T. J., y Tibshirani, R. J. (1986). Generalized Additive Models. *Statistical Science*, *1*(3), 297-310.

Hastie, T. J., y Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall.

Lancaster, P., y Salkauskas, K. (1986). *Curve and Surface Fitting: An Introduction*. Academic Press Inc.

Ruppert, D., Sheater, S. J., y Wand, M. P. (1995). An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of the American Statistical Association*, *90*(432), 1257-1270.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall.

Stone, C. J. (1985). Additive Regression and Other Nonparametric Models. *The Annals of Statistics*, *13*(2), 689-705.

Wand, M. P., y Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R (2ª edición)*. Chapman & Hall.

Wood, S. N. (2023). Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. Paquete de R. Descargado de https://cran.r-project.org/web/packages/mgcv/mgcv.pdf