Grado en Matemáticas

TRABAJO FIN DE ESTUDIOS

Modelos aditivos generalizados en análisis del cambio climático

Tutor: José Luis Pino Mejías

Alumno: Francisco José Lozano Ruiz

Sevilla, Junio de 2024

Índice general

Prólogo

Dedicado a las mujeres de mi vida.
A Chary, que me la dió y me enseñó a vivirla.
A Encarna, que la cuidó como si fuera suya.
A Sebastiana, que la cuida como si fuera suya.
Y a Marina, que le da sentido.
También a mi padre, que espero que se sienta orgulloso allá dónde descanse.

Resumen

El cambio climático es uno de los mayores desafíos a los que nos enfrentamos en la actualidad como sociedad. Ya hemos comenzado a sufrir sus consecuencias y, por tanto, el desarrollo de estrategias que frenen sus efectos y nos permitan adaptarnos a ellos es crucial. Para abordar y entender mejor estos problemas en la mayoría de ocasiones entra en juego el análisis de datos, ya que se trata de una herramienta muy eficaz para este tipo de estudios.

En este trabajo nos proponemos como objetivo el estudiar uno de los métodos comúnmente empleados en estos análisis, los modelos aditivos generalizados. Estos nos permitirán modelar relaciones complejas entre variables de manera flexible, en particular son útiles para capturar relaciones no lineales y efectos de interacción entre variables, algo que es crucial para el análisis exhaustivo del cambio climático.

Abstract

Climate change is one of the greatest challenges we face today as a society. We have already begun to experience its consequences, and therefore, the development of strategies which mitigate its effects and enable us to adapt to them is crucial. In order to tackle and understand better these issues, data analysis is frequently used, as it is a highly effective tool for such studies.

In this project, we aim to study one of the most common methods employed in these analyses, generalized additive models. These models allow us to flexibly model complex relationships between variables. In particular, they are useful for capturing non-linear relationships and interaction effects between variables, which is essential for a comprehensive analysis of climate change.

Índice de figuras

Capítulo 1

Introducción

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

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

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

Comenzaremos el siguiente capítulo introduciendo los Modelos Aditivos y varios resultados básicos sobre la estimación de los MAG, a partir de ellos ya nos podremos adentrar más en profundidad en los resultados necesarios para sus futuras aplicaciones. En particular, hablaremos de cómo elegir los parámetros de los que dependen estos modelos y de las funciones de suavizado (*smoothers*) que aplicaremos luego en la práctica.

Tras el desarrollo del marco teórico, en el capítulo 4 se hace una introducción al cambio climático y a la importancia que tiene su análisis en la actualidad. Los objetivos de este capítulo son tanto concienciar a la población de la gravedad de la situación como estudiar los posibles escenarios futuros y así ser capaces de desarrollar políticas de adaptación y mitigación que nos pertiman actuar ante ellos.

Por último haremos tres aplicaciones de los MAG y los resultados descritos en los capítulos anteriores para distintos fenómenos relacionados con el cambio climático. El capítulo 5 recoge: un estudio de la evolución del clima a rango local, otro sobre la concentración de gases de efecto invernadero y una última aplicación que analiza el incremento del nivel del mar a lo largo de los últimos años. También emplearemos estos modelos para obtener predicciones a futuro.

Capítulo 2

Modelos lineales generalizados

2.1. Introducción

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

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

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

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

Este trabajo tiene como finalidad el definir un conjunto de estrategias y técnicas que proporcionan un ajuste óptimo de la función f, es decir, que asemeje lo mejor posible la relación de las variables al fenómeno que se esté estudiando.

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

[1] El cómo de bueno es un modelo es un concepto que se puede definir de tantas formas como maneras hay de evaluarlos. Veremos varias de ellas a lo largo del trabajo.

2.2. Modelos lineales

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

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

$$\mu = X\beta$$

$$Y = \mu + \epsilon$$

O vectorialmente como:

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

Donde:

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

Observación 2.2.1. Gracias a la definición anterior podemos ver que:
$Y \backsim N(\mu, \sigma^2 I_n)$ ya que:

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

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

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

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

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

Es decir:

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

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

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

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

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

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

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

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

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

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

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

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

y, por tanto:

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

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

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

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

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

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

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

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

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

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

2.3. Modelos lineales generalizados

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

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

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

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

Donde:

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

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

2.3.1. Familia de distribuciones exponenciales

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

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

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

Donde:

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

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

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

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

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

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

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

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

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

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

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

Y su derivada respecto de theta:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.3.2. Ajuste de los modelos lineales generalizados

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

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

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

Y la función de verosimilitud logarítmica:

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

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

en la sección anterior, nos queda:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.3.3. Funciones de enlace canónicas

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

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

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

2.3.4. Residuos

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

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

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

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

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

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

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

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

Donde:

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

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

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

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

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

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

Capítulo 3

Modelos aditivos generalizados

3.1. Introducción

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

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

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

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

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

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

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

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

Donde:

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

3.2. Suavizado univariante

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

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

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

3.2.1. Bases de funciones

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

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

Para ciertos parámetros β_j a determinar.

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

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

Y, por tanto, el modelo 3.2 queda:

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

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

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

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

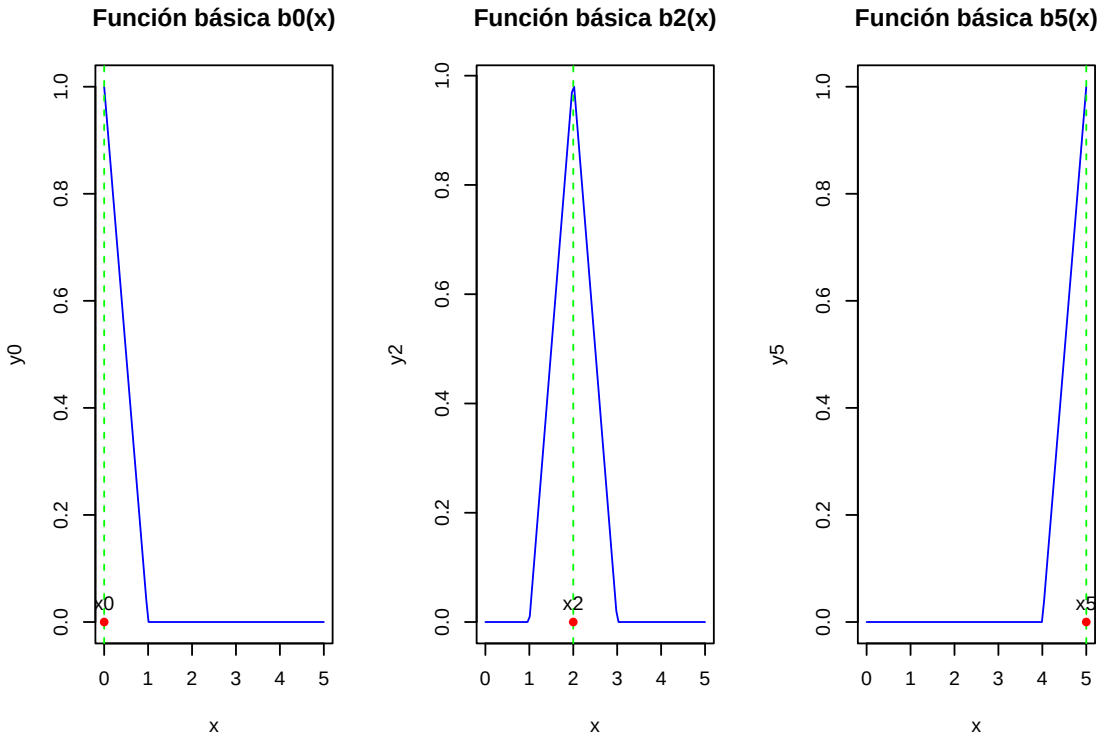
$$b_1(x) = \begin{cases} \frac{x_2^* - x}{x_2^* - x_1^*} & , si \;\; x < x_2^* \\ 0 \quad c.c. \end{cases}$$

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

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

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

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



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

3.2.2. Control del suavizado

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

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

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

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

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

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

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

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

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

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

3.2.3. Elección del parámetro de suavizado

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.3. Modelos aditivos

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

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

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

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

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

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

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

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

Observación 3.3.2. Los principales problemas del modelo aditivo son:

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

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

Modelo aditivo por regresión penalizada por partes

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Ajuste del modelo aditivo por mínimos cuadrados penalizados

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

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

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

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

3.4. Smoothers

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

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

3.4.1. Splines cúbicos

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

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

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

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

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

Splines cúbicos naturales como interpoladores

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

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

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

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

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

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

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

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

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

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

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

Splines cúbicos de suavizado

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

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

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

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

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

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

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

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

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

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

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

3.4.2. Smoothers unidimensionales

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

Splines cúbicos de regresión

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

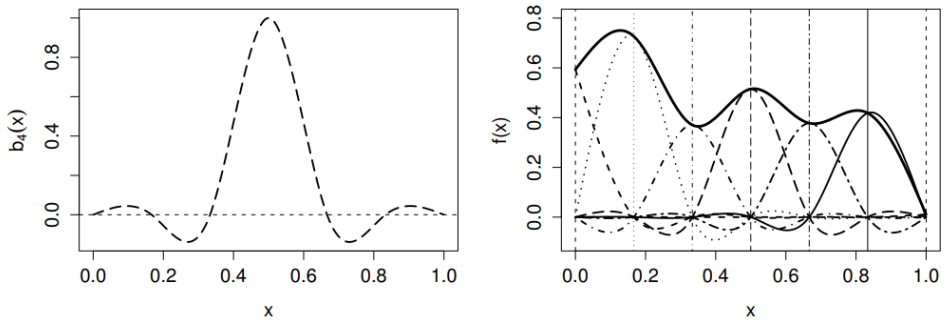


Figura 3.1: Ejemplos de función básica y spline cúbico

Figura 3.1: A la izquierda se ve representada la funcion básica $b_4(x)$ para los splines cúbicos de regresión que se desarrollan en esta sección, esta toma el valor 1 en el nodo 0.5 y 0 en los demás. La cuadrícula de la derecha muestra cómo las funciones básicas se combinan para representar una curva de suavizado. Con menor grosor y con formatos de linea distintos se pueden distinguir las funciones básicas y de forma más llamativa se ve la suma escalada de ellas por los coeficientes β_j. (Wood [2017], p.203).

B-Splines

Veámos ahora otra base para el espacio de funciones de splines de grado p con nodos $x_1 < x_2 < \cdots < x_{k+1}$ arbitrarios. La principal ventaja de la base de B-Splines es que las funciones básicas se definen de forma local, lo que ofrecerá ventajas computacionales frente a las demás bases.

Para definir la base de B-Splines de forma recursiva como en Delicado, denotaremos $M = p + 1$ y consideraremos $2M$ nodos auxiliares τ_i de forma arbitraria de forma que:

$$\tau_1 \leq \tau_2 \leq \cdots \leq \tau_M \leq x_0 \ , \ x_{k+1} \leq \tau_{k+M+1} \leq \cdots \leq \tau_{k+2M}$$

Observemos que los nuevos nodos al escogerse arbitrariamente se podrían tomar como:

$$\tau_1 = \cdots = \tau_M = x_0 \ , \ x_{k+1} = \tau_{k+M+1} = \cdots = \tau_{k+2M}$$

Reescribiremos la nueva particion como: $\tau_{j+M} = x_j \ \forall j = 1, \ldots, k$. Luego, se define la base de B-Splines de forma recursiva como:

$$B_j^{-1}(x) = \begin{cases} 1, \ si \ x_j \leq x < x_{j+1} \\ 0, \ en \ caso \ contrario \end{cases} \quad \forall j = 1, \ldots, k + 2M - 1$$

$$B_j^m(x) = \frac{x - \tau_j}{\tau_{j+m-1} - \tau_j} B_j^{m-1}(x) + \frac{\tau_{j+m} - x}{\tau_{j+m} - \tau_{j+1}} B_{j+1}^{m-1}(x), \quad \forall j = 1, \ldots, k + 2M - m$$

Observación 3.4.3. Estas bases fueron desarrolladas como bases muy estables para la interpolación de splines a gran escala, sin embargo la mayoría de trabajos estadísticos con splines de regresión penalizados de rango menor (es decir, con métodos de suavizado que imponen una reduccion de rango en la matriz asociada al suavizado) deben utilizar métodos numéricos muy deficientes antes de que esta estabilidad sea perceptible. El verdadero interés de la estadística en los B-Splines surge en el desarrollo de los P-Splines.

P-Splines

Los P-Splines se definen como suavizadores que utilizan bases de B-Splines para una partición uniforme de nodos a los que se les aplica una penalización de suavizado para los parámetros β_i. Una forma sencilla de definirlos es utilizando la penalización de la diferencia al cuadrado de los parámetros β_j adyacentes, es decir, dados los nodos $x_1 < \cdots < x_k$ se considera la penalización:

$$\mathcal{P} = \sum_{i=1}^{k-1} (\beta_{i+1} - \beta_i)^2 = \beta^T P^T P \beta$$

Donde:

$$P = \begin{pmatrix} -1 & 1 & 0 & \cdots & \cdots \\ 0 & -1 & 1 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \ddots \end{pmatrix}, \quad P\beta = \begin{pmatrix} \beta_2 - \beta_1 \\ \vdots \\ \beta_k - \beta_{k-1} \end{pmatrix}$$

Luego:

$$\mathcal{P} = \beta \begin{pmatrix} -1 & 1 & 0 & \cdots & \cdots & \cdots \\ -1 & 2 & -1 & 0 & \cdots & \cdots \\ 0 & -1 & 2 & -1 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \ddots \end{pmatrix} \beta$$

Observación 3.4.4. Los P-Splines son fáciles de aplicar y ofrecen una buena flexibilidad a la hora de combinar cualquier orden de penalización con cualquier tamaño de base de los B-Splines. Sin embargo, si se utiliza una partición no uniforme de los nodos, su simplicidad disminuye.

Otra penalización que podemos utilizar es sobre las derivadas de los splines definidos en $[a, b]$. Sean m_1 el orden de la base de B-Splines y m_2 el orden de diferenciación requerido para la penalización:

$$J = \int_a^b f^{m_2}(x)^2 dx = \beta^T S \beta$$

como en 3.4.2. Se debe entonces buscar S, para ello consideraremos una partición ordenada de nodos: $a = x_1 < \cdots < x_{k-m_1+1} = b$ de modo que $h_j = x_{j+1} - x_j \ \ \forall j = 1, \ldots, k-m$ que definirá la base de B-Splines y razonamos de la siguiente forma:

1. Para cada intervalo $[x_j, x_{j+1}]$ generar $p+1$ puntos igualmente espaciados tal que para $p = 0$ se elige el punto central del intervalo y para $p \geq 1$ se deben elegir los extremos x_j y x_{j+1}. Denotaremos el conjunto ordenado de estos puntos por $\mathcal{C}$.
2. Obtener la matriz G aplicando los coeficientes del spline a la m_2-ésima derivada del spline f en los puntos de $\mathcal{C}$.
3. Si $p = 0$, entonces $W = diag(h_j)$.
4. Si $p > 0$, entonces consideremos $P, H \in \mathcal{M}^{(p+1)\times(p+1)}$ con elementos: $P_{i,j} = (\frac{-1+2(i-1)}{p})^j$ y $H_{i,j} = \frac{1+(-1)^{i+j-2}}{i+j-1} \ \ \forall i, j = 1, \ldots, p+1$ y calculemos $\tilde{W} = P^{(-T)} H P^{-1}$. Además, tomar $W = \sum_q W^q$, donde W^q es 0 en todos lados excepto en $W^q_{i+pq-p, j+pq-p} = \frac{h_q}{2} \tilde{W}_{i,j} \ \ \forall i, j = 1, \ldots, p+1$. W es una matriz banda con $2p+1$ diagonales no nulas.
5. La matriz banda de coeficientes de penalización es: $S = G^T W G$.

3.5. Modelos aditivos generalizados

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

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

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

- Y un vector aleatorio n-dimensional, cuyas componentes siguen una distribución de tipo exponencial $y_i \backsim EF(\mu_i, \phi) \ \ \forall i = 1, \ldots, n$.
- X una matriz de orden $n \times p$, $p \leq n$, de columnas $x_j \ \ \forall j = 1, \ldots, p$ con constantes conocidas.

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

Definimos el Modelo Aditivo Generalizado como:

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

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

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

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

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

$$\mathcal{X} = (A \mid X^{[1]} \mid X^{[2]} \mid \cdots)$$

3.5.1. Ajuste del modelo

Seguiremos las directrices dadas en Wood [2017] para dar la estimación de los parámetros β cuando suponemos dado los parámetros de suavizado λ. Partimos notando que el vector de parámetros β contiene a γ y a los vectores de parámetros de suavizado individuales y observemos que una penalización del grado de suavizado global para el modelo viene dada por:

$$\sum_j \lambda_j \beta^T S_j \beta$$

Donde λ_j es un parámetro de suavizado y S_j es por abuso de notación la anterior S_j incrustada como un bloque diagonal en una matriz para que $\lambda_j\beta^T S_j\beta$ sea la penalización para f_j.

Observemos que de este modo el modelo se reduce a un MLG sobreparametrizado de la forma:

$$g(\mu_i) = \mathcal{X}_i\beta, \ \ y_i \backsim EF(\mu_i, \theta_i)$$

El cual se estima maximizando la log-verosimilitud penalizada:

$$l_p(\beta) = l(\beta) - \frac{1}{2\phi}\sum_j \lambda_j \beta^T S_j \beta \tag{3.10}$$

Observación 3.5.1. Aquí, como en 3.2.2, los parámetros de suavizado controlan la compensación entre la calidad de ajuste del modelo y su grado de suavizado.

Ahora, en Wood [2017] se hace notar que esta función objetivo coincide con la de un Modelo Linear Generalizado Mixto (MLGM)[2] y se propone estimarlo por la siguiente variación del 2.3.3.

Definición 3.5.2 (Algoritmo de mínimos cuadrados penalizados reponderados iterativamente)**.**
1. Inicialización: tomar $\hat{\mu_i} = y_i + \delta_i$ y $\hat{\eta_i} = g(\hat{\mu_i})$, donde δ_i suele ser 0 ó una constante que asegure que $\hat{\eta_i}$ sea finito.
2. Calcular: $z_i = g'(\mu_i)\frac{y_i - \hat{\mu_i}}{\alpha(\mu_i)} + \hat{\eta_i}$ y $\omega_i = \frac{\alpha(\hat{\mu}_i)}{g'(\hat{\mu}_i)^2 V(\hat{\mu}_i)}$
3. Encontrar $\hat{\beta}$ el parámetro que minimiza la función objetivo de mínimos cuadrados ponderados penalizados 3.10:

$$||z - \mathcal{X}\beta||^2_W + \sum_j \lambda_j \beta^T S_j \beta$$

y actualizar $\hat{\eta} = \mathcal{X}\hat{\beta}$ y $\hat{\mu_i} = g^{-1}(\hat{\eta_i})$

Donde $||a||^2_W = a^T W a$, $V(\mu)$ es la función de varianza asociada a la distribución de la familia exponencial y $\alpha(\mu_i) = [1 + (y_i - \mu_i)(\frac{V'(\mu_i)}{V(\mu_i)} + \frac{g''(\mu_i)}{g'(\mu_i)})]$.

[2]No nos adentramos en los MLG ya que hace falta su desarrollo para la comprensión del algoritmo.

Capítulo 4

Cambio climático

En la convención de 1992 de las Naciones Unidas sobre el cambio climático se recoge la defición de cambio climático como: " *El cambio de clima atribuido directa o indirectamente a la actividad humana que altera la composición de la atmósfera mundial y que se suma a la variabilidad natural del clima observada durante periodos de tiempo comparables.* "

Ahora bien, para una correcta comprensión de esta definición es necesario dar una noción de clima, ya que existe una confusión generalizada entre los conceptos de clima y tiempo. Por un lado, con tiempo atmosférico nos referimos al estado de la atmósfera en un momento y lugar preciso, este depende de una serie de variables climáticas como la presión, la temperatura, las precipitaciones, etc. Por otro lado, el clima se entiende como las condiciones meteorológicas dadas en un mismo lugar durante un periodo largo de tiempo.

4.1. Consenso científico

Desde la convención de Río de Janeiro antes mencionada se tiene un consenso internacional sobre la existencia del cambio climático, sin embargo todavía no se ha llegado a tal consenso respecto al efecto que ha tenido la humanidad en él, es decir, respecto al Calentamiento Global Antropogénico (CGA). Lo más cercano a ello es el consenso científico.

Para tratar este tema nos referiremos al artículo Myers et al. [2021] de la editorial IOPScience, el cual hace una sintésis de distintas investigaciones sobre el consenso entre expertos sobre el cambio climático y realiza su propio estudio comparándolo con uno de más de 10 años atrás. Este estudio se basa en una encuesta dirigida a la comunidad de ciencias de la Tierra, en el artículo se desarrolla su metodología y se desglosan todos los resultados diferenciando entre distintas características de los encuestados (área en la que están especializados, número de publicaciones científicas, número de publicaciones relevantes...), pero nosotros nos centraremos únicamente en sus conclusiones.

Lo primero que se destaca de la investigación es que, en comparación a otra del 2009, el número de científicos que no aceptan el impacto humano en el cambio climático se ha reducido a más de la mitad, pasa del 20 % a tan solo el 9 %, es decir, ahora alrededor de un 91 % de los encuestados están de acuerdo con la acción que tiene el hombre en el calentamiento global. Además, si distinguimos los grados de experiencia de los encuestados

por el número de publicaciones relevantes sobre el cambio climático en los últimos años, resulta que aquellos con más de 15 publicaciones relevantes aceptan el efecto humano en un 100 % y aquellos con más de 10 en un 98.5 % (1 profesional de entre los 67 que abarca el grupo).

También es interesante destacar el siguiente par de conclusiones que saca el artículo: el área de los expertos encuestados con un menor porcentaje de acuerdo con el calentamiento global antropogénico es el de la geología económica (7 % menos que la media) y aquellos profesionales con más años de experiencia (sin depender del número de publicaciones científicas) en el campo también tienden a tener una menor aceptación. Podemos observar los resultados de forma más detallada en las siguientes gráficas.

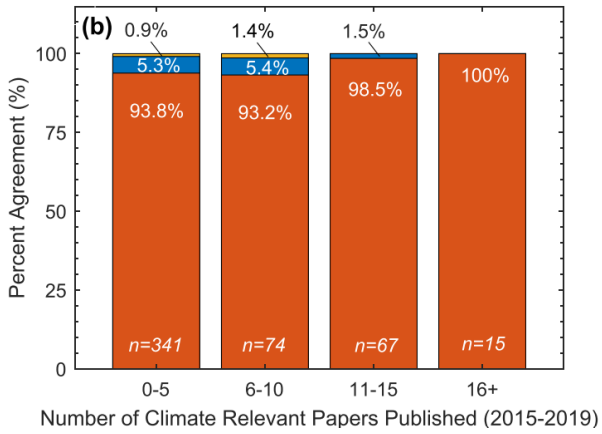


Figura 4.1: Porcentaje de aceptación respecto del número de publicaciones relevantes en los últimos años.

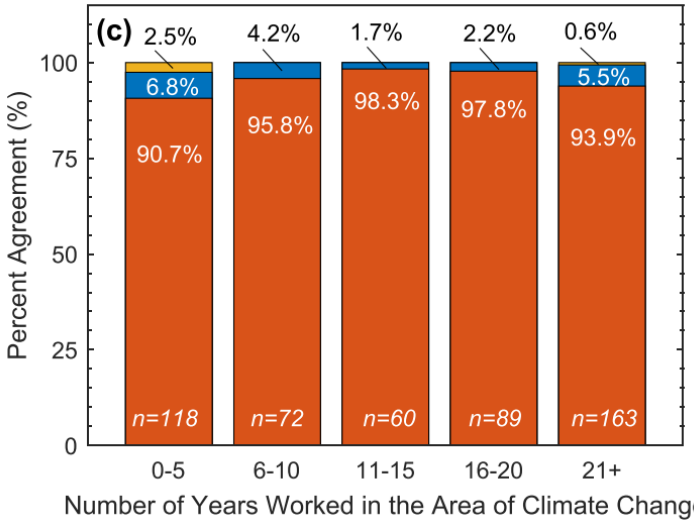


Figura 4.2: Porcentaje de aceptación respecto de los años trabajados en el ámbito del cambio climático.

El 2.5 % que aparece en el segundo diagrama se corresponde con aquellos que no solo niegan el CGA sino que niegan por completo la existencia del cambio climático.

4.2. Controversia

Sin embargo, aunque exista un consenso internacional sobre la existencia del cambio climático y uno científico sobre el efecto que tiene la humanidad en él, aún hay personas y organizaciones que lo niegan. Este hecho conlleva a que las instituciones implicadas en la lucha del cambio climático se vean envueltas en una crisis de la desinformación, además de la crisis medioambiental que se proponen combatir. Estos contenidos negacionistas no solo socavan el apoyo público mediante información y anuncios dudosos, sino que también erosionan la confianza en estas instituciones.

Varios estudios de ciencias sociales han analizado estas posturas como formas de negacionismo, pseudociencia e incluso propaganda. La coalición de la Acción Contra la Desinformación Climática (CAAD sus siglas en inglés) realizó el informe CAA [2023] , resumido en el artículo de Limb [2023] de EuroNews, en el que se puede intuir quiénes están detrás de estas fuentes de desinformación y el por qué de sus intereses en ellas. En este artículo se recoge que, ahora más que nunca, es importante que las sociedades tengan una visión compartida sobre el cambio climático que se base en datos y en la integridad de la información. Sin embargo, la desinformación es una piedra en el camino para este objetivo, pues además de poner al cambio climático en el punto de mira de teorías conspirativas y de ser una causa para la división social, perjudica la implementación y el debate de nuevas políticas climáticas.

En el informe antes mencionado se detallan tres casos en los que la desinformación y los bulos son perjuciales para la lucha del cambio climático:

- Por un lado, nombra una serie de páginas webs y editoriales que publican desinformación sobre el clima y se lucran a través de bolsas de anuncios. Esto significa que muchas marcas posicionadas en contra de, por ejemplo, la emisión de gases contaminantes aparezcan junto a estos artículos y de forma indirecta permitan su monetización.
- Por otro lado, expone un caso asociado a los medios de comunicación estatales rusos. El CAAD afirma que estos medios utilizan campañas de desinformación con mensajes incoherentes y falsos para reforzar planes de influencia contra países Occidentales y del sur Global.
- Por último habla sobre las grandes compañías petroleras, las cuales invierten millones de dólares en publicidad con el objetivo de lavar la imagen de su marca. Tienen presupuestos inmensos destinados al desarrollo de anuncios en los que suelen alardear de sus inversiones en energías renovables, pero el análisis del CAAD muestra que estas empresas sólo aportaron el 1

No obstante, estos ejemplos no son más que la punta del iceberg de una gran crisis de la información, ya que existen muchas más entidades e individuos que sacan provecho de una cuestión tan importante como esta y a la par que perjudican a los organismos que pretenden combatirla.

Citemos a continuación una serie de argumentos utilizados por los negacionistas:

- Sólo se tienen datos desde el 1960, el clima comenzó a calentarse antes de la revolución industrial.

- El dióxido de carbono se quedará en la atmósfera durante décadas o siglos.
- Se han producido otros cambios climáticos a lo largo de la historia.
- Se achacan catástrofes al calentamiento global aunque no hay pruebas para vincularlas.
- Si hay una ola de frío, ¿Cómo va a haber calentamiento global? (Dicho por Donald Trump en 2018).
- El cambio climático tiene efectos positivos como obtener mejores cosechas de cereales en los países que son menos cálidos.
- No hay calentamiento global desde 1998.
- No están aumentando los fenómenos meteorológicos extremos.

4.3. Resumen del cambio climático en la actualidad

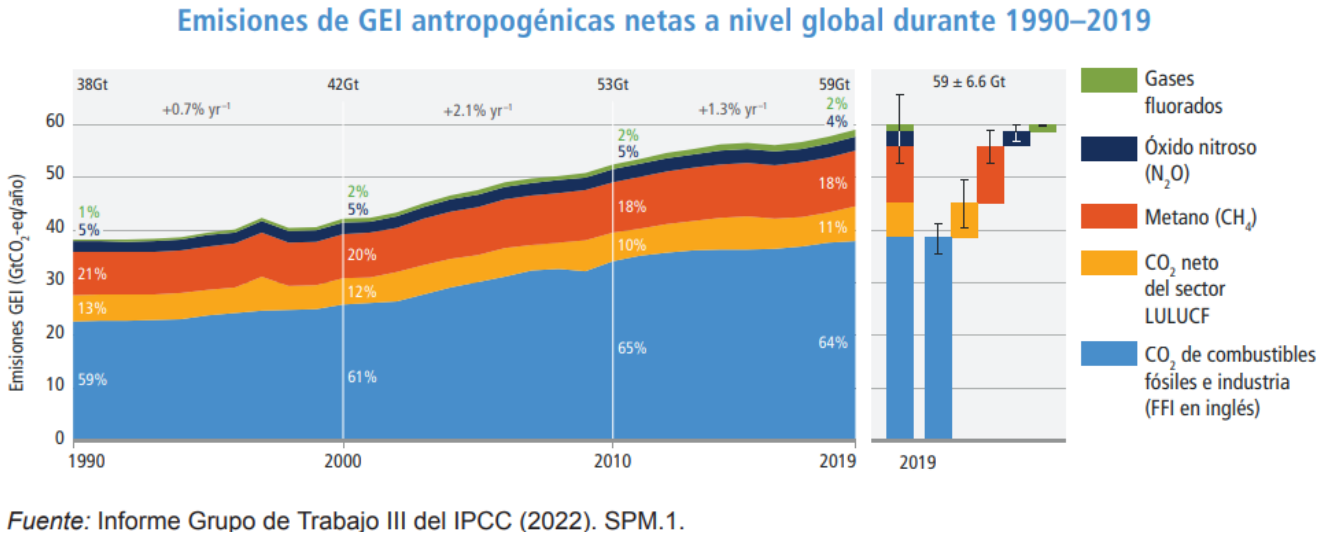
Nos basaremos en el sexto informe del Grupo Intergubernamental de Expertos sobre Cambio Climático (más conocido por sus siglas en inglés, IPCC). El IPCC es una entidad científica creada en 1988 por la Organización Meteorológica Mundial y el Programa de Naciones Unidas para el Medio Ambiente y tiene como principal propósito el proporcionar información objetiva, clara, equilibrada y neutral del estado actual de conocimientos sobre el cambio climático. Lo haremos refiriéndonos a los resúmenes en español realizados por la Oficina Española de Cambio Climático, los cuales se pueden encontrar en su web Climático.

El cambio climático constituye un fenómeno global, tanto por sus causas como por sus efectos, y requiere una respuesta multilateral basada en la colaboración de todos los países, ya no solo a nivel internacional, sino también a nivel regional.

El primer grupo de trabajo abre el apartado sobre la situación actual del cambio climático con la siguiente aserción: " *Es indiscutible que las actividades humanas están causando un cambio climático, haciento que los eventos extremos sean más frecuentes y severos*". Tras ello, exponen las siguientes observaciones:

- La concentración de CO_2 en la atmósfera es la más alta en los últimos 2 millones de años.
- El incremento de la subida del mar en el último siglo es mayor que al registrado en cualquier siglo anterior en los últimos 3000 años.
- La superficie de hielo marino ártico está en su nivel más bajo en los últimos 1000 años.
- La temperatura media en la última década (2011-2020) fue aproximadamente 1.09°C superior a la temperatura media entre 1850 y 1900.

Lo siguiente que nos interesa notar está desarrollado por el grupo III de trabajo del IPCC, el cual se encarga del informe de mitigación. Hacen una evaluación actual de las tendencias de las emisiones antropogénicas de gases contaminantes a nivel global y por sectores. Por un lado, afirman que las emisiones de gases de efecto invernadero (GEI) han seguido aumentando a nivel global entre 2010 y 2019. Aunque en la siguiente gráfica podemos observar que, en comparación a la década anterior, se ha reducido el aumento por año de las emisiones netas en un 0.8 %, el total aumenta ya hasta las 59Gt (Gigatoneladas). [1] [2]



Fuente: Informe Grupo de Trabajo III del IPCC (2022). SPM.1.

Figura 4.3: Emisión de GEI desde 1990 hasta 2019

Además de ello, comparan las emisiones netas acumuladas de CO_2 entre 2010 y 2019 con las que se pueden emitir a partir de 2020 para limitar el calentamiento global a 1.5°C y a 2°C, estas representarían $\frac{4}{5}$ de las emisiones para la primera y aproximadamente $\frac{1}{3}$ para la segunda. La Unión Europea comenta en Europea cuáles pueden ser algunas de las consecuencias de este incremento de temperaturas:

- Aumento de la mortalidad.
- Reducción de la productividad.
- Cambios en la distribución geográfica de las zonas climáticas que alteren el ecosistema de especies vegetales y animales.
- Influencia en la fenología, el comportamiento y los ciclos de vida de las especies. Esto puede hacer que aumente el número de plagas y de especies invasoras, así como la incidencia de algunas enfermedades humanas.
- Aumento de la evaporación del agua, lo que, unido a la falta de precipitaciones, aumenta el riesgo de sequías graves.

[1] LULUCF son las siglas en inglés de "uso de la tierra, cambio de uso de la tierra y selvicultura".

[2] GtCO_2-eq es una medida en toneladas de la huella de carbono,es decir, de la totalidad de la emisión de gases de efecto invernadero.

Por otro lado, presentamos el siguiente gráfico realizado por la Oficina Española de Cambio Climático en el que se dividen las emisiones antropógenicas de gases de efecto invernadero según el sector de proveniencia. Se puede apreciar una reducción de emisiones de CO_2 procedentes de los combustibles fósiles y de la industria, aunque no fue suficiente para compensar el aumento de emisiones del resto de sectores debido al incremento de los niveles de actividad global.

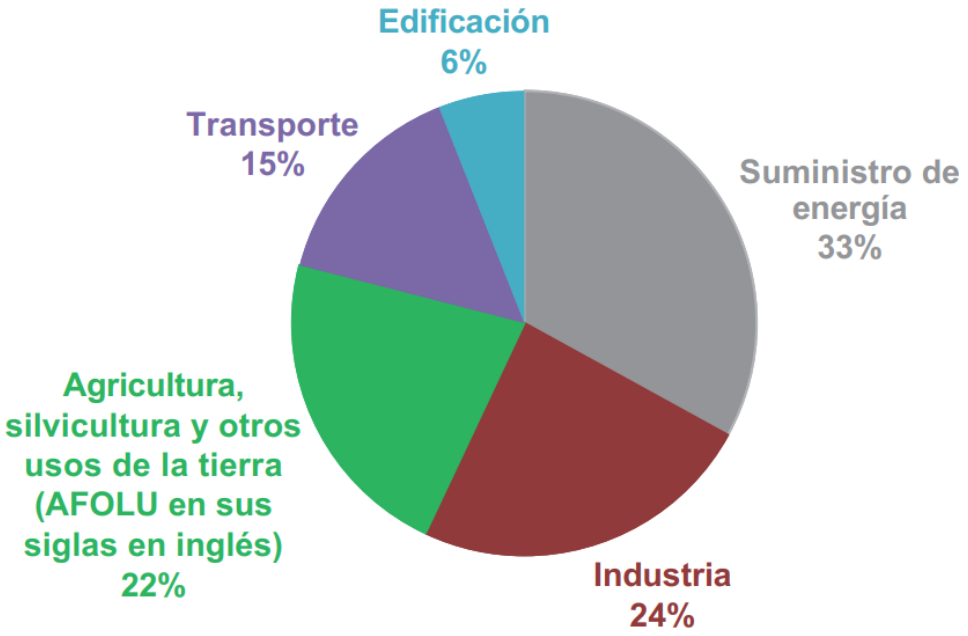


Figura 4.4: Emisiones globales de GEI por sectores 2019

4.4. Acciones de adaptación y mitigación

El Panel Intergubernamental sobre Cambio Climático (IPCC [2022]) propone varias soluciones de adaptación y mitigación para abordar el cambio climático, las cuales se pueden dividir en medidas para reducir las emisiones de gases de efecto invernadero y estrategias para adaptarse a los cambios ya inevitables. Se pueden ver recogidas en el artículo Boehm and Schumer [2023].

Soluciones de Mitigación

- Transición energética: abandonar los combustibles fósiles y aumentar la inversión en energías renovables como solar y eólica.
- Eficiencia energética: mejorar la eficiencia energética en todos los sectores, incluyendo edificios, transporte e industria.
- Reducción de emisiones en la Agricultura: implementar prácticas agrícolas sostenibles como la agroforestería y la agricultura de conservación para reducir las emisiones.

- Conservación y restauración de ecosistemas: proteger y restaurar bosques, humedales y otros ecosistemas que secuestran carbono.
- Reducción de residuos y economía circular: minimizar los residuos y fomentar el reciclaje y la reutilización de materiales para reducir las emisiones de GEI relacionadas con la producción y el consumo.

Soluciones de Adaptación

- Infraestructuras resilientes: construir y renovar infraestructuras para que sean más resistentes a eventos climáticos extremos, como inundaciones y tormentas.
- Gestión del agua: implementar sistemas de gestión del agua que aseguren el suministro durante periodos de sequía y gestionen el exceso durante inundaciones.
- Adaptación de la agricultura: cultivos resistentes al clima y prácticas agrícolas adaptativas que pueden soportar condiciones cambiantes.
- Seguros y mecanismos financieros: desarrollar mecanismos financieros como seguros contra desastres naturales y fondos de contingencia en caso de eventos climáticos extremos.

Capítulo 5

Aplicación de los MAG al análisis del cambio climático

En esta sección nos proponemos aplicar el contexto teórico visto hasta ahora sobre los modelos aditivos generalizados al análisis del cambio climático. Para ello principalmente utilizaremos el paquete de R: *mgcv* (siglas en inglés de "Vehículo de Computación para MAG Mixtos"), en particular haremos uso de su funcion *gam*, la cual permite ajustar modelos aditivos generalizados, entre otros tipos de modelos, mediante splines de regresión penalizados (u smoothers similares) donde los parámetros de suavizado pueden ser estimados por distintos métodos, por ejemplo por: mínima validación cruzada generalizada, mínimo AIC, máxima verosimilitud o por REML, siglas en inglés de máxima verosimilitud restringida, que es la opción por defecto.

Además del método de estimación, la función *gam* también admite otras entradas que sirven para indicar qué familia de distribuciones exponenciales se utiliza, si las observaciones toman distintos pesos, el método de optimización numérica utilizado, otros parámetros de control de estos métodos para el caso de que los habituales no converjan, etc.

Dividiremos las aplicaciones prácticas de los MAG en tres partes: la primera se centra en modelar la temperatura media mensual según una serie de variables climáticas y ver cómo ha variado con los años, en la siguiente veremos cómo han evolucionado a lo largo del tiempo las concentraciones de gases de efecto invernadero en la atmósfera, por último estudiaremos la media variacional del nivel del mar respecto del año 1993.

La lectura y transformación de los datos no la mostraremos en esta sección, se encuentra en A y el código que genera las representaciones gráficas aparece en B.

5.1. Modelización de la temperatura media mensual

5.1.1. Descripción de los datos

Para esta primera aplicación de los modelos aditivos generalizados utilizaremos datos de variables climáticas proporcionados por la Agencia Estatal de Meteorología española (AEMET), para ello utilizaremos la libreria "climaemet" Pizarro et al. [2021]:

```
#install.packages('climaemet')
library(climaemet)
```

Como hemos dicho antes, el conjunto de datos sobre el que trabajaremos a lo largo de esta sección está formado por variables climáticas, estas se definen como elementos que caracterizan el tiempo atmosférico y que interactuan entre sí en la troposfera. Aunque son elementos relacionados con el campo de la meteorología, su estudio a largo plazo, fundamenta las bases científicas de la climatología. En particular, el conjunto de datos mensuales que nos proporciona la anterior librería contiene más de 40 variables climáticas, por comodidad y para una mejor interpretación de los modelos nos quedaremos con las variables que representen la temperatura media mensual, la humedad relativa, la media mensual de precipitaciones, la presión media y la velocidad media del viento.

Esta información provienen de una base de datos *open source* que ofrece la AEMET, se puede leer más sobre ella en: https://opendata.aemet.es/centrodedescargas/inicio. En particular utilizaremos las mediciones tomadas por la estación situada en el aeropuerto de Sevilla.

Hagamos la primera visualización de los datos observando su estructura y resumen:

```
str(Clima)
```

```
## tibble [768 x 7] (S3: tbl_df/tbl/data.frame)
##  $ Año  : num [1:768] 1960 1960 1960 1960 1960 1960 1960 1960 1960 1960 ...
##  $ Mes  : Factor w/ 12 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ HR   : num [1:768] 88 85 87 76 73 73 48 45 47 75 ...
##  $ PresM: num [1:768] 1011 1013 1015 1013 1012 ...
##  $ Prec : num [1:768] 81.3 205.2 108.7 18.4 45 ...
##  $ WMed : num [1:768] 9 12 16 12 12 12 11 12 13 15 ...
##  $ TMedM: num [1:768] 10.3 13 14.4 17.4 21 25.9 27.2 25.9 24.3 17.1 ...
```

```
summary(Clima)
```

```
##       Año             Mes             HR            PresM           Prec       
##  Min.   :1960   1      : 64   Min.   :31.00   Min.   :1004   Min.   :  0.00  
##  1st Qu.:1976   2      : 64   1st Qu.:51.00   1st Qu.:1011   1st Qu.:  2.00  
##  Median :1992   3      : 64   Median :61.00   Median :1013   Median : 26.40  
##  Mean   :1992   4      : 64   Mean   :60.74   Mean   :1014   Mean   : 45.81  
##  3rd Qu.:2007   5      : 64   3rd Qu.:71.00   3rd Qu.:1016   3rd Qu.: 65.25  
##  Max.   :2023   6      : 64   Max.   :90.00   Max.   :1028   Max.   :361.10  
##                 (Other):384   NA's   :9       NA's   :5                      
##       WMed            TMedM       
##  Min.   : 5.00   Min.   : 8.40  
##  1st Qu.: 9.00   1st Qu.:13.50  
##  Median :11.00   Median :18.10  
##  Mean   :11.09   Mean   :18.93  
##  3rd Qu.:12.00   3rd Qu.:24.65  
##  Max.   :22.00   Max.   :30.70  
##  NA's   :17      NA's   :5
```

- HR: la humedad relativa media es un valor porcentual de la cantidad de vapor de agua presente en el aire con respecto a la máxima posible para unas condiciones dadas de presión y temperatura.
- PresM: la presión media mensual al nivel de la estación medida en hectopascales (hPa).
- Prec: la precipitación total mensual medida en milímetros.
- WMed: la velocidad media del aire se mide en metros por segundo.
- TMedM: la temperatura media mensual viene dada en grados centígrados.

5.1.2. Descripción del modelo

Tomaremos a la variable *TMedM* como variable de respuesta y al resto como variables explicativas. Antes de definir el modelo debemos notar que la variable *Año* no se ha definido como variable categórica, como sí se hizo para la variable *Mes*, sino que se define como variable numérica para luego poder tener una mejor representación de los resultados obtenidos. Además, si se hubiera definido de tal forma, resultaría que la mayoría de factores son no significativos. Dicho esto definimos el modelo como:

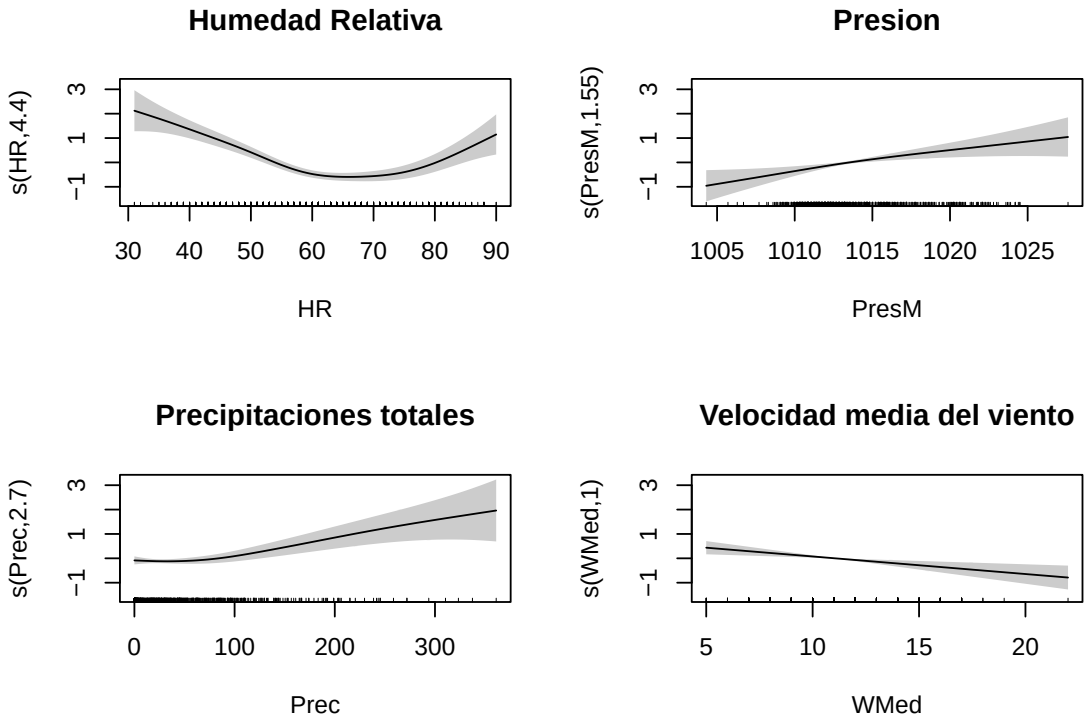
```
#install.packages('mgcv')
library(mgcv)
```

```
mag1 <- gam(TMedM~ s(HR)+s(PresM)+s(Prec)+s(WMed)+Año+Mes,data = Clima)
summary(mag1)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## TMedM ~ s(HR) + s(PresM) + s(Prec) + s(WMed) + Año + Mes
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -55.026379   5.890706  -9.341  < 2e-16 ***
## Año           0.032945   0.002917  11.294  < 2e-16 ***
## Mes2          1.923150   0.232581   8.269 6.56e-16 ***
## Mes3          4.537181   0.272233  16.667  < 2e-16 ***
## Mes4          6.970016   0.318541  21.881  < 2e-16 ***
## Mes5         10.287254   0.340623  30.201  < 2e-16 ***
## Mes6         13.898675   0.351116  39.584  < 2e-16 ***
## Mes7         16.644447   0.378536  43.971  < 2e-16 ***
## Mes8         16.817040   0.373449  45.032  < 2e-16 ***
## Mes9         14.188786   0.329066  43.118  < 2e-16 ***
## Mes10         9.765255   0.283822  34.406  < 2e-16 ***
```

```
## Mes11         4.157867   0.238799  17.412  < 2e-16 ***
## Mes12         0.798865   0.218239   3.661  0.00027 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##            edf Ref.df      F  p-value
## s(HR)    4.396  5.468 19.901  < 2e-16 ***
## s(PresM) 1.551  1.945  6.211 0.001852 **
## s(Prec)  2.704  3.410  6.083 0.000262 ***
## s(WMed)  1.000  1.000 10.396 0.001320 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.962   Deviance explained = 96.3%
## GCV = 1.4701  Scale est. = 1.4251    n = 740
```

Como podemos ver en el resumen del modelo, se toma por defecto que la variable dependiente sigue la distribución normal y que la función de enlace es la identidad. Se tiene que el modelo es capaz de explicar el 96.3 % de la varianza con $R^2_{adj} = 0.962$ y que todas las variables predictoras son significativas. Observemos qué efecto tienen las variables predictoras sobre la temperatura media mensual:



De las gráficas de la derecha se puede interpretar que los efectos de *PresM* y *WMed* sobre la temperatura media mensual son lineales. Ajustamos entonces un nuevo modelo aditivo generalizado del mismo modo que antes pero imponiendo que el efecto de estas variables sea lineal:

```r
mag2 <- gam(TMedM~ s(HR)+PresM+s(Prec)+WMed+Año+Mes,data = Clima)
summary(mag2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## TMedM ~ s(HR) + PresM + s(Prec) + WMed + Año + Mes
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.388e+02  2.690e+01  -5.158 3.22e-07 ***
## PresM        8.417e-02  2.400e-02   3.507 0.000481 ***
## WMed        -7.201e-02  2.235e-02  -3.222 0.001332 **
## Año          3.253e-02  2.886e-03  11.272  < 2e-16 ***
## Mes2         1.943e+00  2.315e-01   8.394 2.50e-16 ***
## Mes3         4.559e+00  2.713e-01  16.804  < 2e-16 ***
## Mes4         6.972e+00  3.184e-01  21.899  < 2e-16 ***
## Mes5         1.029e+01  3.402e-01  30.245  < 2e-16 ***
## Mes6         1.391e+01  3.507e-01  39.648  < 2e-16 ***
## Mes7         1.664e+01  3.782e-01  44.011  < 2e-16 ***
## Mes8         1.681e+01  3.730e-01  45.064  < 2e-16 ***
## Mes9         1.420e+01  3.285e-01  43.242  < 2e-16 ***
## Mes10        9.791e+00  2.827e-01  34.635  < 2e-16 ***
## Mes11        4.188e+00  2.366e-01  17.702  < 2e-16 ***
## Mes12        8.095e-01  2.181e-01   3.711 0.000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df      F  p-value
## s(HR)   4.469  5.551 20.250  < 2e-16 ***
## s(Prec) 2.610  3.294  5.994 0.000342 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.962   Deviance explained = 96.3%
## GCV = 1.4708  Scale est. = 1.4269    n = 740
```

Podemos ver que tanto la desviación explicada como la estimación del error por validación cruzada coinciden con las del modelo anterior, sin embargo utilizaremos la función *anova* para dar una prueba de razón de verosimilitud que determine si el modelo más complejo mejora significativamente el ajuste.

```
anova(mag1,mag2,test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: TMedM ~ s(HR) + s(PresM) + s(Prec) + s(WMed) + Año + Mes
## Model 2: TMedM ~ s(HR) + PresM + s(Prec) + WMed + Año + Mes
##   Resid. Df Resid. Dev       Df Deviance      F Pr(>F)
## 1    715.18     1022.3
## 2    716.15     1024.4 -0.97706  -2.1327 1.5316 0.2158
```

Como el p-valor resultante del contraste de hipótesis es > 0.05, no tenemos evidencias significativas como para rechazar la hipótesis nula, es decir, se acepta que ambos modelos tienen el mismo ajuste.

También es posible compararlos mediante otros criterios, por ejemplo el AIC (Akaike Information Criterion) que tiene en cuenta el número de parámetros a estimar y el valor objetivo de la función de log-verosimilitud:

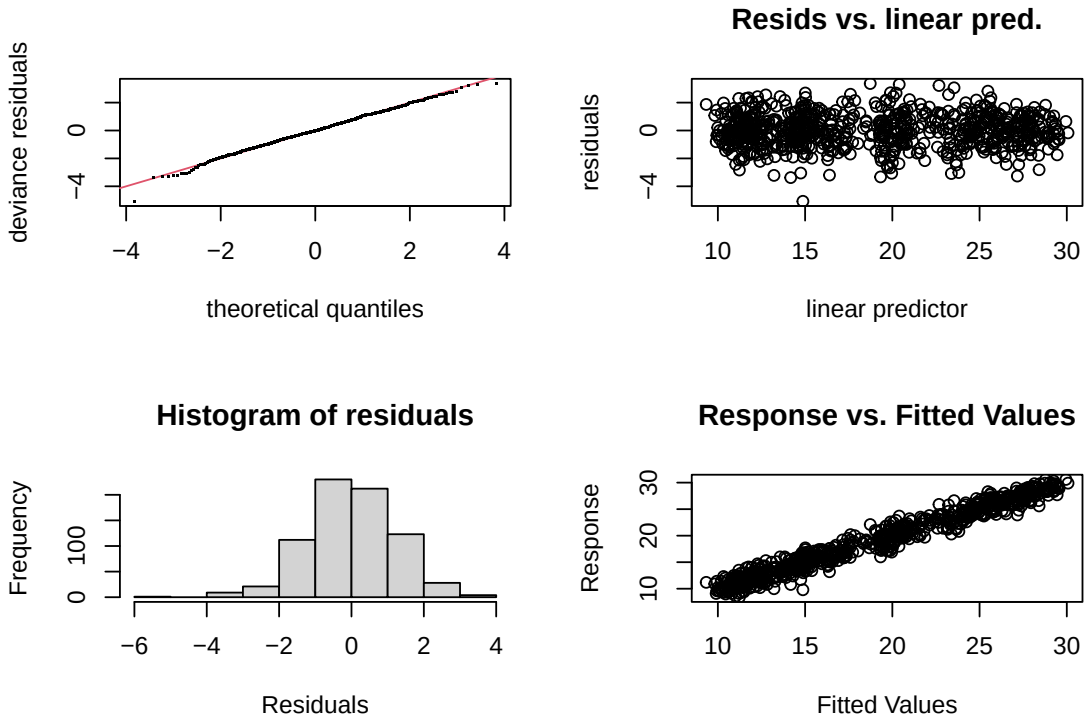
```
AIC(mag1,mag2)
```

```
##            df      AIC
## mag1 23.65122 2386.466
## mag2 23.07915 2386.864
```

En este caso son casi idénticos, aunque el primer modelo tiene menor AIC.

Para obtener más información sobre el modelo planteado se utiliza la siguiente rutina de diagnósticos que nos proporciona información y gráficos útiles para evaluar la calidad del ajuste del modelo.

```
gam.check(mag1)
```



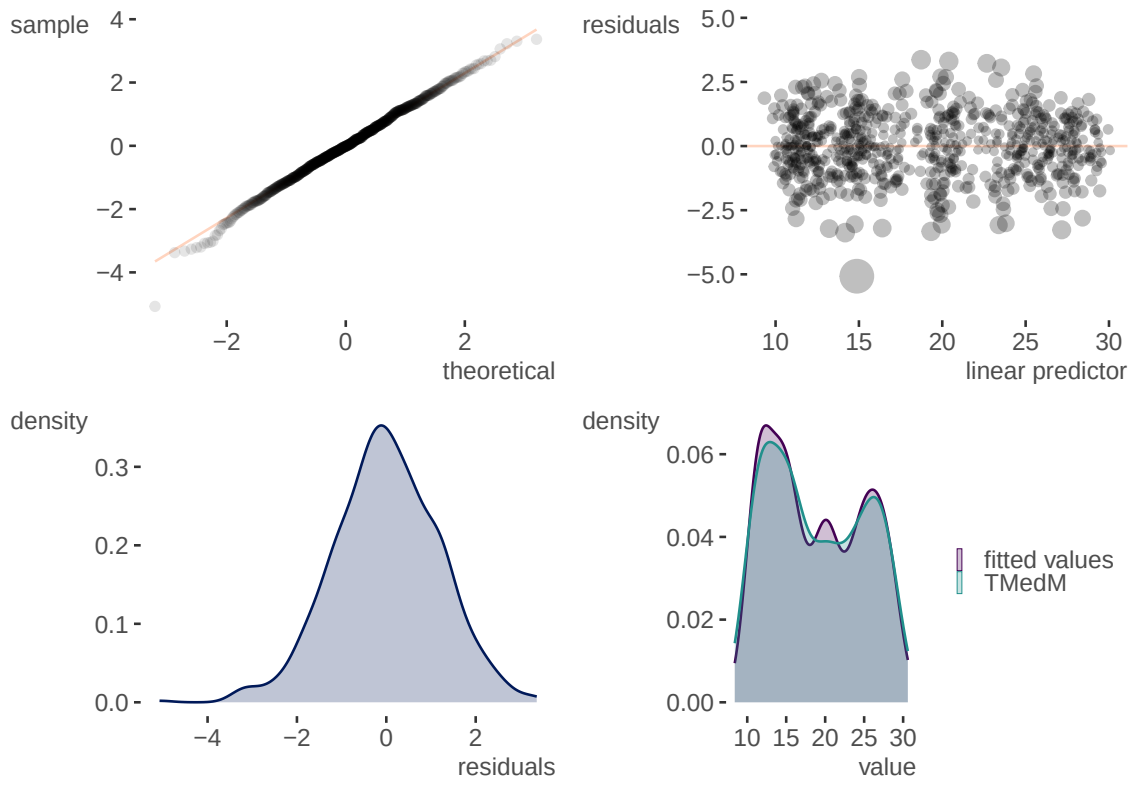

```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 10 iterations.
## The RMS GCV score gradient at convergence was 1.49324e-07 .
## The Hessian was positive definite.
## Model rank =  49 / 49
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##            k'  edf k-index p-value
## s(HR)    9.00 4.40    0.93   0.040 *
## s(PresM) 9.00 1.55    0.92   0.015 *
## s(Prec)  9.00 2.70    0.94   0.095 .
## s(WMed)  9.00 1.00    0.87  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Por un lado, la salida por consola nos informa de que se obtiene la convergencia por optimización del GCV y que el modelo es de rango completo. Los p-valores que aparecen se corresponden a los tests de residuos aleatorios correspondientes para cada predictor, en este caso todos excepto el asociado a *Prec* son < 0.05, lo que indica que se necesitaría una base de funciones de mayor dimensión para la función de suavizado asociada a esta variable. No obstante, no se obtiene ninguna mejoría significativa del modelo con este razonamiento así que nos ahorraremos el incluirlo. Por otro lado, observemos qué representa cada una de las gráficas generadas y cuál sería el caso ideal para cada una de ellas:

- Q-Q Plot: compara la distribución de los residuos con una distribución normal. Lo ideal es que los puntos se alineen aproximadamente en una línea recta.
- Resids vs. linear pred: representa los residuos frente el predictor lineal, ayuda a verificar si los residuos se distribuyen aleatoriamente, que sería lo ideal.
- Histograma de residuos: se trata de un histograma de los residuos, en este caso lo ideal es que muestre una distribución aproximadamente normal, centrada en cero, esto indicaría que los residuos no presentan sesgos significativos.
- Response vs. Fitted Values: representa los valores observados frente a los valores ajustados, lo ideal sería que los puntos resultantes se aconglomerasen en torno a la recta $x = y$.

Por lo general este modelo se comporta de manera correcta, ya que se aproxima mucho a los casos ideales de cada gráfica. Podemos utilizar la librería *visibly* de Clark para hacer esta representación de forma más clara:

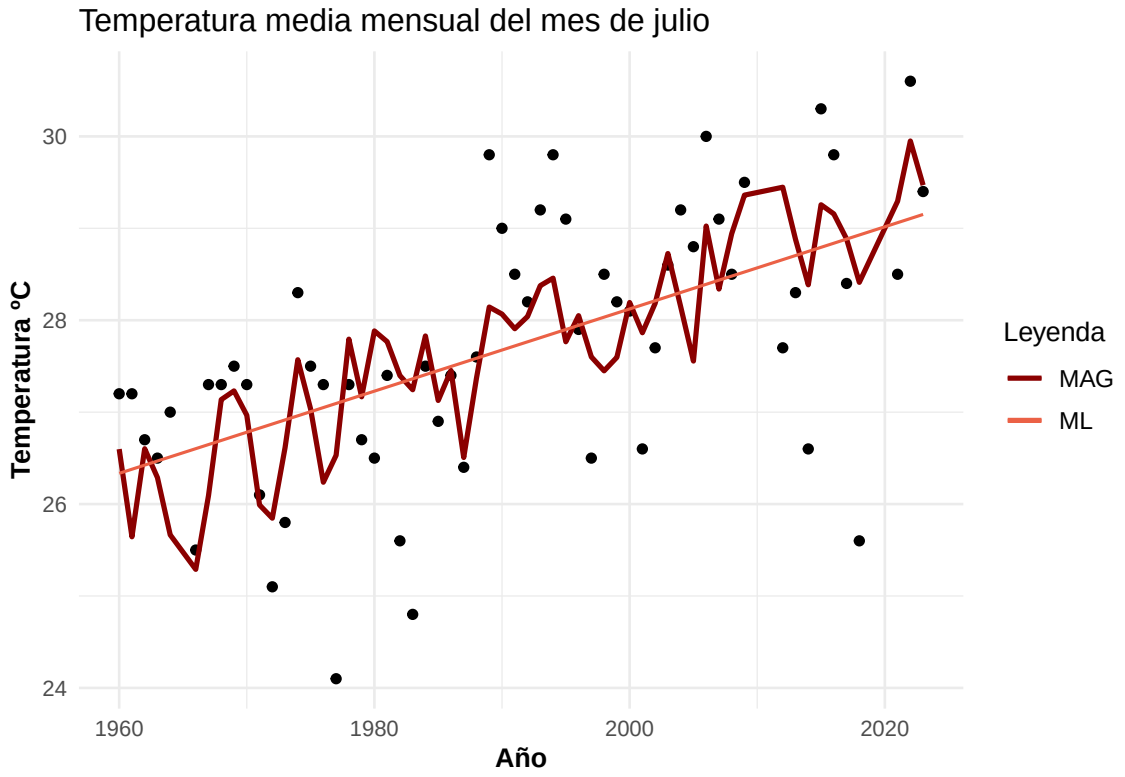
```
#install.packages('visibly')
library(visibly)
plot_gam_check(mag1)
```



5.1.3. Visualización de resultados

Una vez generado el modelo y comprobado que se tiene una buena bondad de ajuste, veremos cómo ajusta los datos con el objetivo de identificar si se ha producido un cambio

significativo en la temperatura media mensual a lo largo de los años. Para evitar las variaciones entre estaciones nos centraremos en la representación de los datos para un mes fijo, en este caso julio.



Con esta gráfica es posible apreciar claramente como la tendencia de la temperatura media en el mes de Julio presenta un aumento constante desde la década de los 60s en la estación meteorológica del aeropuerto de San Pablo. Hemos introducido la recta proporcionada por el modelo lineal para los datos de los meses de Julio para tener una mejor apreciación de tal incremento.

5.2. Modelización de gases de efecto invernadero

5.2.1. Descripción de los datos

En esta sección ajustaremos modelos aditivos generalizados con el objetivo de estudiar la concentración atmosférica de gases de efecto invernadero, en particular analizaremos las medias mensuales globales de las concentraciones de dióxido de carbono (CO_2), metano (CH_4) y óxido nitroso (N_2O). Para ello consideraremos los datos proporcionados por United Nations Environment Programme (UNEP), para el CO_2 se obtienen en https://wesr.unep.org/climate/essential-climate-variables-ecv/atmospheric-co2-concentration y para los dos siguientes en https://wesr.unep.org/climate/essential-climate-variables-ecv/atmospheric-ch4-n2o-sf6-concentration.

Ya hablamos sobre las emisiones de gases contaminantes en 4.3 pero no se llegó a definir en qué consistían. Estos gases son capaces de absorber y emitir radiación dentro

del espectro inflarrojo, por tanto son capaces de retener el calor del Sol, lo que permite que el clima terrestre sea habitable para la humanidad. Sin embargo, desde el inicio de la revolución industrial, la actividad humana ha producido un desequilibrio en los niveles de concentración de estos gases en la atmósfera. En particular estudiaremos las concentraciones de los tres tipos de gases antes mencionados por ser los que se emiten en mayor cantidad o por ser las más potentes (en términos de contribución al efecto invernadero). Por ejemplo, las emisiones de CO_2 se corresponden aproximadamente con tres cuartas partes del total de emisiones de GEI, sin embargo el CH_4 y el N_2O representan una parte mucho menor que el dióxido de carbono pero por unidad son mucho más potentes como gases de efecto invernadero. Christina.

Para facilitar la lectura de los datos se ha transformado el archivo excel proporcionado por la fuente antes mecionada a *.xlsx* y se han utilizado las librerías *readxl* y *lubridate* para leerlo y obtener las variables temporales *Año* y *Mes*, respectivamente.

```
#install.packages('readxl')
library(readxl)
#install.packages('lubridate')
library(lubridate)
```

- CO_2:

```
str(CO2)
```

```
## tibble [794 x 4] (S3: tbl_df/tbl/data.frame)
##  $ Año  : num [1:794] 1958 1958 1958 1958 1958 ...
##  $ Mes  : Factor w/ 12 levels "1","2","3","4",..: 3 4 5 6 7 8 9 10 11 12 ...
##  $ Tmes : num [1:794] 316 317 318 317 316 ...
##  $ Trend: num [1:794] NA NA NA NA NA NA NA NA NA NA ...
```

```
summary(CO2)
```

```
##       Año             Mes           Tmes           Trend
##  Min.   :1958   3      : 67   Min.   :312.4   Min.   :316.0
##  1st Qu.:1974   4      : 67   1st Qu.:330.4   1st Qu.:331.1
##  Median :1991   1      : 66   Median :355.0   Median :355.7
##  Mean   :1991   2      : 66   Mean   :359.0   Mean   :359.3
##  3rd Qu.:2007   5      : 66   3rd Qu.:384.6   3rd Qu.:384.0
##  Max.   :2024   6      : 66   Max.   :426.6   Max.   :421.1
##                 (Other):396                   NA's   :729
```

De este modo nos queda un data frame con 794 observaciones, correspondientes a los meses desde marzo del 1958 hasta abril de 2024, y con las variables *Año*, *Mes*, *Tmes* que se corresponde con la concentración media de CO_2 a nivel global medida en partes por millón (ppm) y *Trend* que es la media anual de las anteriores. El tener 1 ppm de un gas en un medio significa que existe una molécula de ese gas por cada millón de moléculas de aire.

- CH_4:

```
str(CH4)
```

```
## tibble [487 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Año  : num [1:487] 1983 1983 1983 1983 1983 ...
##  $ Mes  : Factor w/ 12 levels "1","2","3","4",..: 7 8 9 10 11 12 1 2 3 4 ...
##  $ Trend: num [1:487] 1635 1636 1636 1637 1638 ...
```

```
summary(CH4)
```

```
##       Año            Mes          Trend
##  Min.   :1983   1      : 41   Min.   :1635
##  1st Qu.:1993   7      : 41   1st Qu.:1737
##  Median :2003   8      : 41   Median :1775
##  Mean   :2003   9      : 41   Mean   :1778
##  3rd Qu.:2013   10     : 41   3rd Qu.:1816
##  Max.   :2024   11     : 41   Max.   :1928
##                 (Other):241
```

En este caso disponemos de 487 observaciones correspondientes a los meses entre 1983 y 2024, las variables de tiempo *Año* y *Mes* y la variable *Trend* la cual representa la media mensual de concentración de metano a nivel global medida en *parts per billion* (ppb). La notación es en inglés, en español se correspondría a partes por miles de millones.

- N_2O:

```
str(N2O)
```

```
## tibble [277 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Año  : num [1:277] 2001 2001 2001 2001 2001 ...
##  $ Mes  : Factor w/ 12 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Trend: num [1:277] 316 316 316 316 316 ...
```

```
summary(N2O)
```

```
##       Año            Mes          Trend
##  Min.   :2001   1      : 24   Min.   :316.0
##  1st Qu.:2006   2      : 23   1st Qu.:320.0
##  Median :2012   3      : 23   Median :325.1
##  Mean   :2012   4      : 23   Mean   :325.6
##  3rd Qu.:2018   5      : 23   3rd Qu.:330.7
##  Max.   :2024   6      : 23   Max.   :337.3
##                 (Other):138
```

Para el caso del óxido nitroso sólo disponemos datos desde el 2001, por lo que obtenemos un conjunto de 277 observaciones para las variables *Año*, *Mes* y *Trend* que representa la media mensual de concentración de N_2O a nivel global medida en *parts per billion* (ppb).

Solo con los resúmenes de los datos para los tres gases, teniendo en cuenta las medidas en las que vienen dados, ya se puede ver la gran diferencia que hay entre sus proporciones en la atmósfera.

5.2.2. Descripción de los modelos

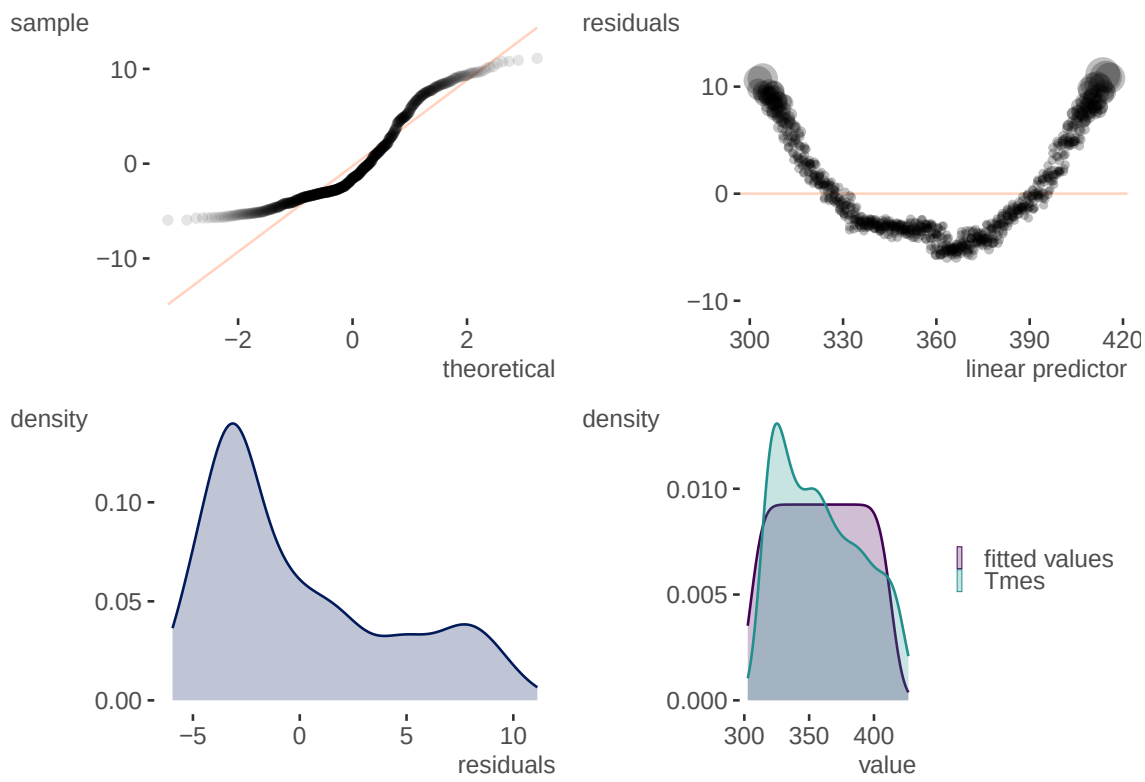
Partiremos definiendo un modelo lineal para los datos de CO_2 que tenga como variable de respuesta la media mesual global de la concentración de este gas medida en ppm y como predictoras las variables *Año* y *Mes*:

```
magCO2 <- gam(Tmes ~ Año + Mes, data = CO2)
summary(magCO2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Tmes ~ Año + Mes
##
## Parametric coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -2.897e+03  1.637e+01 -176.948  < 2e-16 ***
## Año          1.635e+00  8.215e-03  199.022  < 2e-16 ***
## Mes2         7.917e-01  7.697e-01    1.029 0.303997
## Mes3         1.770e+00  7.668e-01    2.308 0.021232 *
## Mes4         3.071e+00  7.668e-01    4.004 6.81e-05 ***
## Mes5         3.486e+00  7.697e-01    4.530 6.84e-06 ***
## Mes6         2.925e+00  7.697e-01    3.800 0.000156 ***
## Mes7         1.390e+00  7.697e-01    1.806 0.071250 .
## Mes8        -6.210e-01  7.697e-01   -0.807 0.420035
## Mes9        -2.164e+00  7.697e-01   -2.812 0.005046 **
## Mes10       -2.111e+00  7.697e-01   -2.743 0.006228 **
## Mes11       -7.661e-01  7.697e-01   -0.995 0.319867
## Mes12        5.563e-01  7.697e-01    0.723 0.470079
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =   0.98   Deviance explained = 98.1%
## GCV = 19.874  Scale est. = 19.549    n = 794
```

Podemos observar que representa un 98.1 % de la desviación explicada, por lo que en principio no parece un mal ajuste, comprobémoslo con el diagnóstico de residuos como se hizo en el apartado anterior:

```
plot_gam_check(magCO2)
```



Gracias a estos gráfios se puede ver cómo el modelo falla en varios aspectos:

- En primer lugar, el Q-Q plot no se adapta a la recta, es decir, la distribución de los residuos no es normal.
- En la gráfica de arriba a la derecha se puede ver como los residuos toman un patrón claro respecto de los predictores, lo que implica que la hipótesis de homocedasticidad de los residuos no es cierta.
- En la de abajo a la izquierda se ve facilmente que no es una distribución normal centrada en el 0, lo que confirma lo observado en el Q-Q plot.
- Por último, en el gráfico Response vs. Fitted también se puede intuir la falta de homocedasticidad para los residuos.

Luego, aunque el modelo representase un gran porcentaje de la desviación explicada, presenta errores significativos en el análisis de los residuos por lo que inferimos que el modelo no es adecuado. Para definir un modelo que se ajuste mejor nos referiremos a dichas gráficas. En primer lugar, que la gráfica de residuos frente predictores se asemeje a una función cuadrática nos indica que puede existir una relación no lineal entre la varible de respuesta y las predictoras (que es tal y como se definió el modelo), por lo que será conveniente añadir funciones de suavizado al modelo. Además, también puede implicar que la familia de distribuciones exponenciales para la variable de respuesta que estemos utilizando, la normal en este caso, no sea la adecuada. Podemos utilizar el test de normalidad univariante Shapiro-Wilk para comprobarlo:

```r
shapiro.test(CO2$Tmes)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CO2$Tmes
## W = 0.9407, p-value < 2.2e-16
```

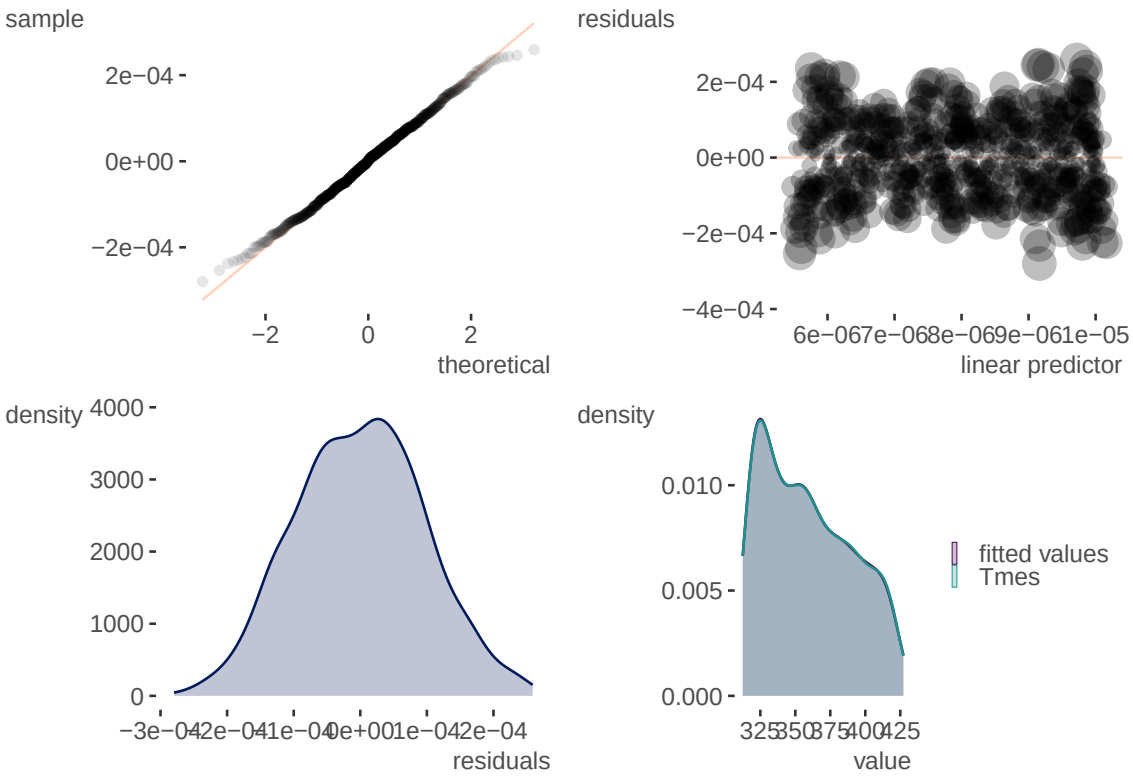
Como el p-valor es muy cercano a 0, se rechaza la hipótesis nula de normalidad de la muestra. Lo que haremos entonces será razonar que como los datos tratados son niveles de concentraciones positivas, quizás nos interese utizar la distribución *Gamma* o la *Inverse Gaussian*. Para determinar cuál de las dos es más conveniente compararemos los modelos definidos para ambas familias con sus respectivas funciones de enlace.

```r
magCO2b <- gam(Tmes ~ s(Año) + Mes, data = CO2,
               family = inverse.gaussian(link = "1/mu^2"))
summary(magCO2b)
```

```
##
## Family: inverse.gaussian
## Link function: 1/mu^2
##
## Formula:
## Tmes ~ s(Año) + Mes
##
## Parametric coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  7.962e-06  3.493e-09 2279.492  < 2e-16 ***
## Mes2        -3.316e-08  4.911e-09   -6.754 2.83e-11 ***
## Mes3        -6.782e-08  4.890e-09  -13.869  < 2e-16 ***
## Mes4        -1.217e-07  4.877e-09  -24.956  < 2e-16 ***
## Mes5        -1.462e-07  4.902e-09  -29.820  < 2e-16 ***
## Mes6        -1.229e-07  4.908e-09  -25.040  < 2e-16 ***
## Mes7        -5.866e-08  4.924e-09  -11.913  < 2e-16 ***
## Mes8         2.682e-08  4.945e-09    5.423 7.86e-08 ***
## Mes9         9.343e-08  4.962e-09   18.828  < 2e-16 ***
## Mes10        9.112e-08  4.962e-09   18.364  < 2e-16 ***
## Mes11        3.304e-08  4.947e-09    6.679 4.58e-11 ***
## Mes12       -2.339e-08  4.933e-09   -4.742 2.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##          edf Ref.df      F p-value
## s(Año) 8.947  8.999 196755  <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =      1   Deviance explained =  100%
## GCV = 9.7546e-09  Scale est. = 9.4978e-09  n = 794
```

```
plot_gam_check(magCO2b)
```

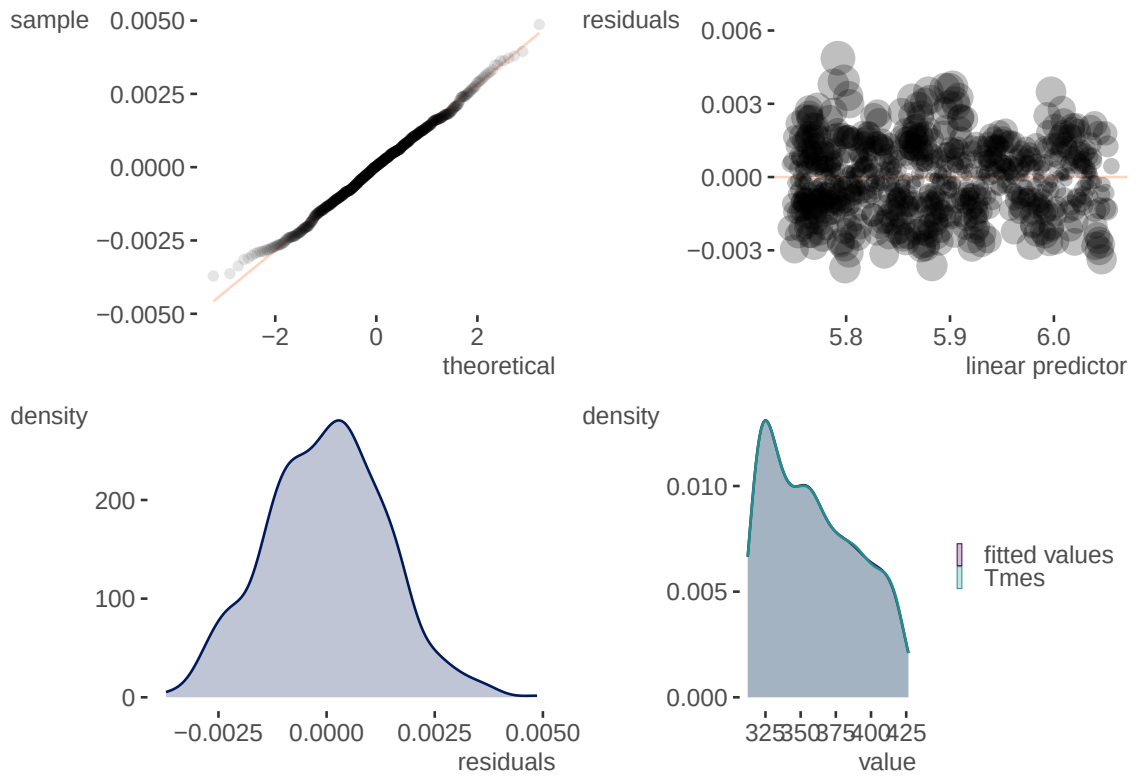



```
magCO2c <- gam(Tmes ~ s(Año) + Mes, data = CO2, family = Gamma(link = "log"))
summary(magCO2c)
```

```
##
## Family: Gamma
## Link function: log
##
## Formula:
## Tmes ~ s(Año) + Mes
##
## Parametric coefficients:
##               Estimate Std. Error   t value Pr(>|t|)
## (Intercept)  5.8777562  0.0001721 34160.995  < 2e-16 ***
## Mes2         0.0021993  0.0002432     9.042  < 2e-16 ***
## Mes3         0.0045205  0.0002424    18.652  < 2e-16 ***
## Mes4         0.0081229  0.0002424    33.516  < 2e-16 ***
## Mes5         0.0097221  0.0002433    39.956  < 2e-16 ***
## Mes6         0.0081689  0.0002433    33.572  < 2e-16 ***
```

```
## Mes7          0.0039223  0.0002433    16.120  < 2e-16 ***
## Mes8         -0.0017190  0.0002433    -7.065 3.59e-12 ***
## Mes9         -0.0061035  0.0002433   -25.084  < 2e-16 ***
## Mes10        -0.0059939  0.0002433   -24.634  < 2e-16 ***
## Mes11        -0.0022295  0.0002433    -9.163  < 2e-16 ***
## Mes12         0.0014633  0.0002433     6.014 2.79e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##          edf Ref.df      F p-value
## s(Año) 8.965      9 341186  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =      1   Deviance explained =  100%
## GCV = 2.0052e-06  Scale est. = 1.9524e-06  n = 794
```

```
plot_gam_check(magCO2c)
```



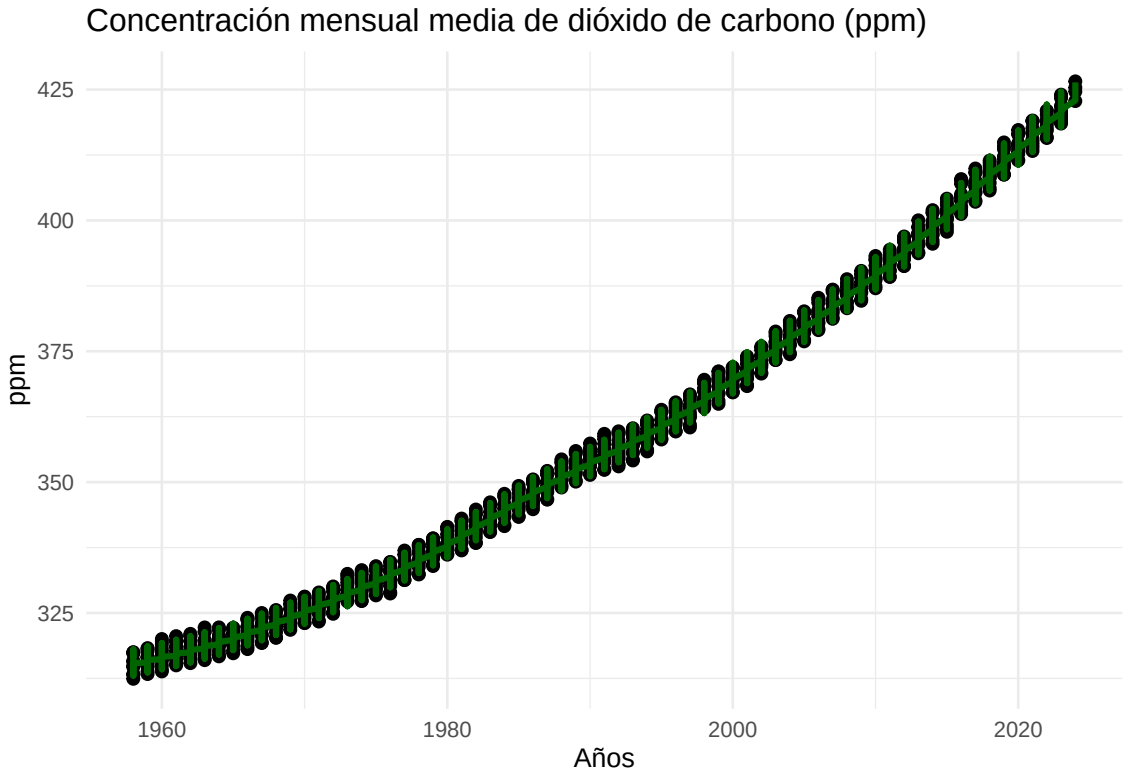
Vemos ahora como las gráficas para ambos diagnósticos ofrecen resultados mucho mejores y que incluso los R^2_{adj} llegan a 1 para ambos modelos. Comparémoslos:

```
##                      AIC          GCV
## Inv. Gauss (b) 1635.459 9.754596e-09
## Gamma (c)      1174.703 2.005172e-06
```

Por un lado, el modelo c tiene un AIC significativamente menor que el modelo b. Esto sugiere que el modelo c es mejor en términos de bondad del ajuste, teniendo en cuenta la penalización de la complejidad. Por el otro lado, el modelo b tiene un GCV significativamente menor que el modelo c, por lo que tedrá una mejor capacidad predictiva. Con cuál quedarnos ya dependerá del objetivo que nos propongamos, para dar un buen ajuste de los datos preferiremos el que utiliza la familia *Gamma* y para predecir nuevas observaciones preferiremos el modelo con la familia *Inverse Gaussian*. Apliquemos estos dos casos.

5.2.3. Visualización de resultados

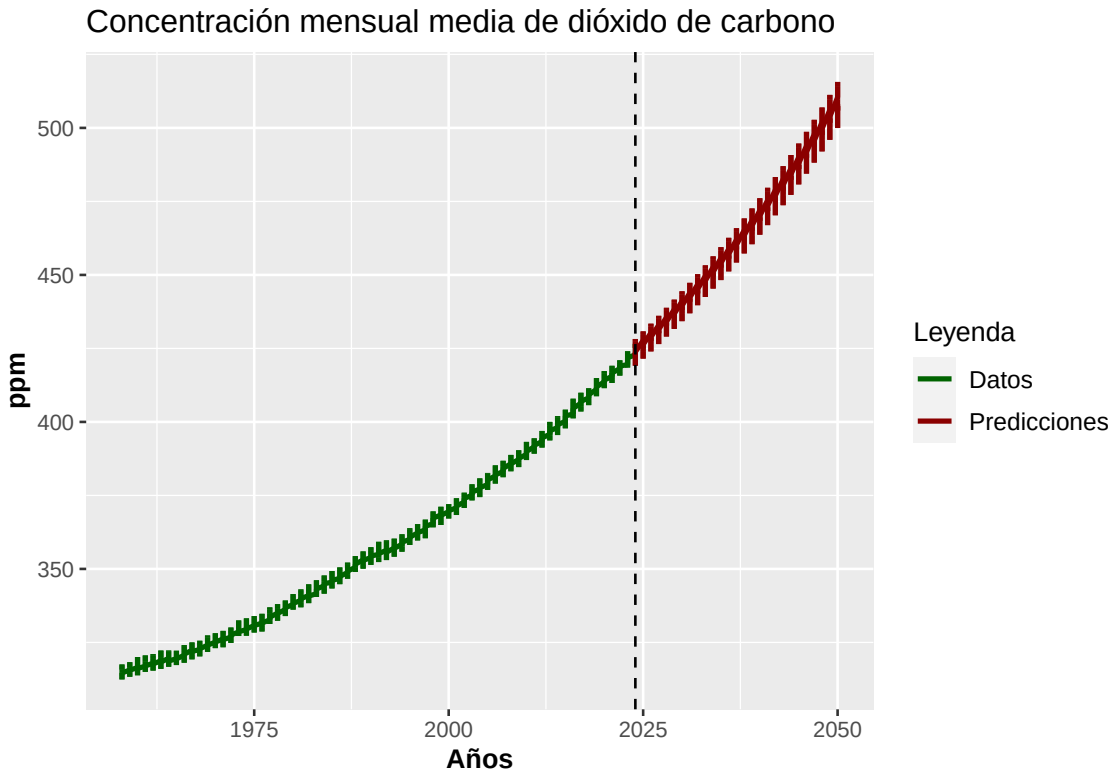
Ajuste de la concentración del dióxido de carbono
Como hemos indicado, el MAG que utiliza la distribución *Gamma* como familia de distribución exponencial proporciona una mejor bondad de ajuste teniendo en cuenta la complejidad del modelo, así que utilizaremos ese para representar el ajuste de los datos. Se debe tener presente que sobre las predicciones se debe invertir la función de enlace utilizada, en este caso la función logarítmica.



Se pueden observar claramente las oscilaciones interestacionales, como se indica en https://climate.nasa.gov/en-espanol/signos-vitales/dioxido-de-carbono/?intent=111, esto se debe a que en la primevera las plantas absorben una mayor cantidad de CO_2 para alimentar su crecimiento. A pesar de ello, se aprecia que el incremento en los niveles de concentración de dióxido de carbono en la atmósfera es de más de 100 ppm en los últimos 60 años.

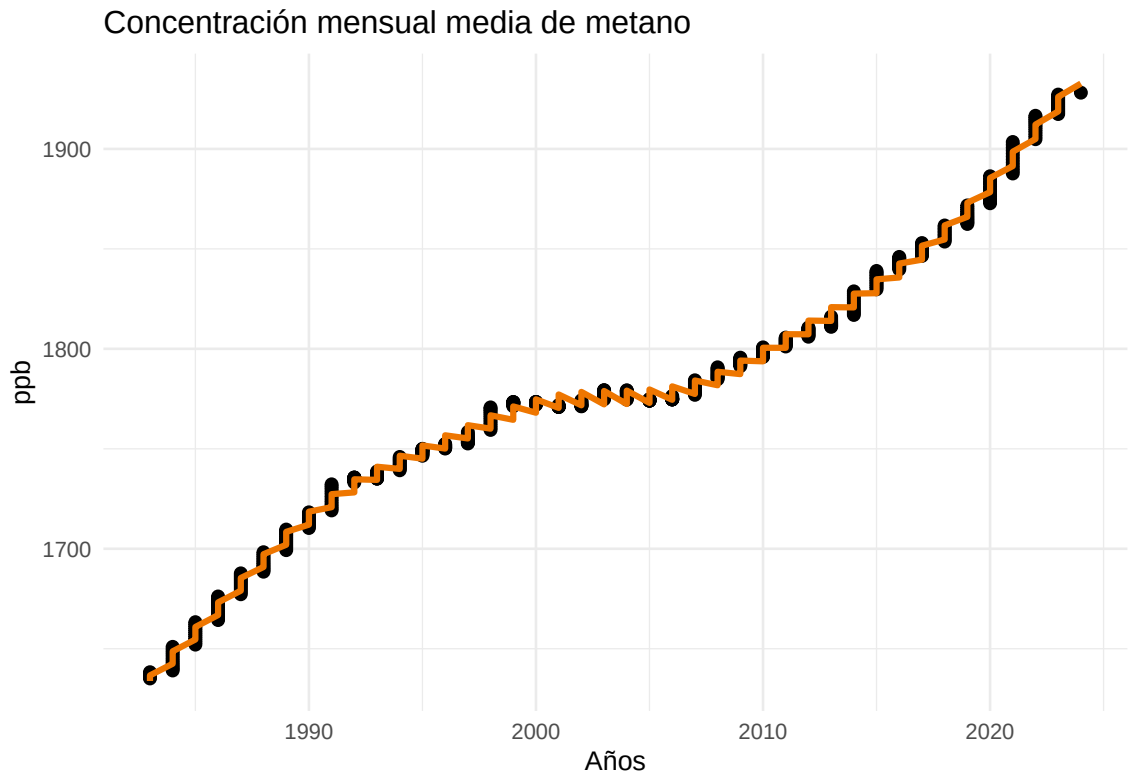
Predicciones de la concentración del dióxido de carbono

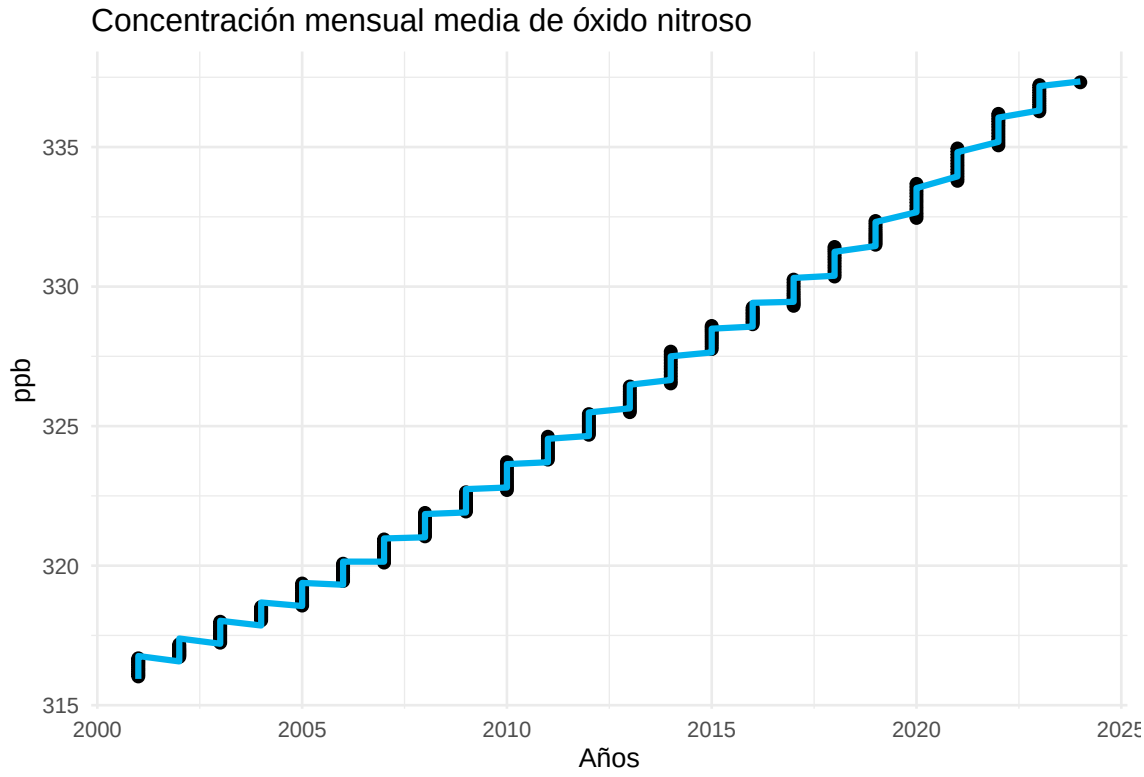
Para representar las predicciones aplicaremos el modelo que utilizaba la distribución *Inverse Gaussian*, pues tenía un menor valor del error estimado por validación cruzada generalizada.



Obviamente esta es una predicción, que tan solo tiene en cuenta el paso del tiempo y los niveles de concentración medidos hasta ahora. Para poder hacer predicciones más exactas se necesitarían datos relativos a las emisiones de CO_2, al clima, al crecimiento de la población y crecimiento económico y se deberían tener en cuenta políticas sobre energías, economía y tecnología.

Seguimos el mismo razonamiento para los conjuntos de datos que contienen las concentraciones de CH_4 y N_2O. Obtenemos la siguientes representaciones de los ajustes para cada modelo:





Se puede apreciar en la primera gráfica que la media mensual de concentración atmosférica de metano ha aumentado en más de 200 ppb desde 1983 y que la correspondiente al N_2O ha aumentado en más de 20 ppb desde el 2001. Comparado con el incremento que se vio anteriormente para el CO_2 parece poco llamativo, pero como se comentó al principio de la sección, se debe tener en cuenta que estos dos gases captan mucha mas

radiación que el primero, lo que hace que su aumento, por poco que sea, también genere preocupación.

5.3. Modelización del aumento del nivel del mar

Ya comentamos en la sección 4.3 que la subida del nivel del mar dada en el último siglo es relevante comparada con la de cualquier siglo anterior. Como se argumenta en Craig [2024] , si este aumento tan pronunciado se progonla a medio-largo plazo pueden darse grandes consecuencias tales como el incremento de la frecuencia y la importancia de las inundaciones en zonas costeras, el aumento de amenazas por fenómenos climáticos extremos como huracanes y grandes tormentas, la erosión del suelo y las costas que puede implicar la pérdida del hábitat de peces, pájaros y plantas, etc. Esto conlleva a que esta sea una de las causas del cambio climático con mayor interés de estudio. Sin embargo, para poder realizar un análisis aceptable de los datos asociados a este hecho primero se debe distinguir si se quiere hacer a nivel global o a nivel local y se necesita disponer de información relacionada con múltiples factores como: la salinidad del agua, la geología del terreno, el deshielo de los polos, las oscilaciones oceánicas, eventos climáticos extremos que puedan ocurrir o hayan ocurrido...

En nuestro caso no disponemos de tanta información así que nos propondremos un objetivo más simple como es el definir un modelo aditivo generalizado que tenga como variable de respuesta la media mensual global del nivel del mar (GMSL por sus siglas en inglés) y como variables predictoras utilizaremos la media mensual de la temperatura de la superficie marítima global, la media de concentración atmosférica de CO_2 (la que utilizamos en la sección anterior como variable dependiente) y los datos temporales de meses y años.

5.3.1. Descripción de los datos

Como acabamos de comentar, las medidas de concentraciones de CO_2 que utilizaremos son las mismas que en la sección anterior así que no entraremos en más detalles para esos datos.

Con respecto a las medidas del GMSL utilizamos la misma fuente de datos que para los gases de efecto invernadero: la UNEP https://wesr.unep.org/climate/essential-climate-variables/sea-level-rise, lo único que debemos tener en cuenta es que se ha modificado el excel de cierta forma para generar las columnas *Year* y *Month*, correspondientes a la fecha en la que se obtuvo la medición, y hemos creado la columna *GMSL* que indica la diferencia en milímetros del nivel del mar con respecto a la media en 1993. Antes de ese año los datos se corresponden con las medidas reconstruidas por la UNEP y a partir de él con la media de una serie de distintas medidas satelitales. El resumen de los datos asociados al nivel del mar es el siguiente:

```
##       Año             Mes            GMSL        
##  Min.   :1900   1      :123   Min.   :-160.10  
##  1st Qu.:1930   2      :123   1st Qu.:-123.20  
##  Median :1961   3      :123   Median : -67.65
```

```
##  Mean   :1961   4       :123   Mean   : -63.39
##  3rd Qu.:1992   5       :123   3rd Qu.: -17.02
##  Max.   :2022   (Other):860   Max.   :  79.31
##  NA's   :1      NA's    :  1   NA's   :2
```

Por otro lado, para obtener los datos respectivos a la temperatura media global nos referiremos a los datos ofrecidos por la NASA en https://data.giss.nasa.gov/gistemp/. Su resumen viene dado por:

```
##       Año             Mes             Temp
##  Min.   :1880   Min.   : 1.00   Min.   :-0.81000
##  1st Qu.:1916   1st Qu.: 3.75   1st Qu.:-0.22000
##  Median :1952   Median : 6.50   Median :-0.03000
##  Mean   :1952   Mean   : 6.50   Mean   : 0.07082
##  3rd Qu.:1988   3rd Qu.: 9.25   3rd Qu.: 0.29250
##  Max.   :2024   Max.   :12.00   Max.   : 1.48000
##  NA's   :12                     NA's   :20
```

Tras haber cargado los tres conjuntos de datos necesarios para aplicar el modelo, los unimos en un solo data frame con observaciones desde 1959 hasta 2014:

```
str(Sea)
```

```
## tibble [672 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Año : num [1:672] 1959 1959 1959 1959 1959 ...
##  $ Mes : Factor w/ 12 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ GMSL: num [1:672] -65.4 -68.7 -70.8 -70.1 -69.7 -68.3 -68 -65.6 -66.6 -65 ...
##  $ Temp: num [1:672] 0.08 0.07 0.18 0.16 0.04 0.03 0.03 -0.01 -0.06 -0.07 ...
##  $ CO2 : num [1:672] 316 316 317 318 318 ...
```

```
summary(Sea)
```

```
##       Año             Mes           GMSL              Temp
##  Min.   :1959   1       : 56   Min.   :-80.700   Min.   :-0.3500
##  1st Qu.:1973   2       : 56   1st Qu.:-48.675   1st Qu.: 0.0400
##  Median :1986   3       : 56   Median :-24.250   Median : 0.2650
##  Mean   :1986   4       : 56   Mean   :-22.486   Mean   : 0.2803
##  3rd Qu.:2000   5       : 56   3rd Qu.:  1.978   3rd Qu.: 0.5200
##  Max.   :2014   6       : 56   Max.   : 47.987   Max.   : 1.0200
##                 (Other):336
##       CO2
##  Min.   :313.3
##  1st Qu.:328.3
##  Median :348.7
##  Mean   :350.9
##  3rd Qu.:370.8
##  Max.   :402.0
##
```

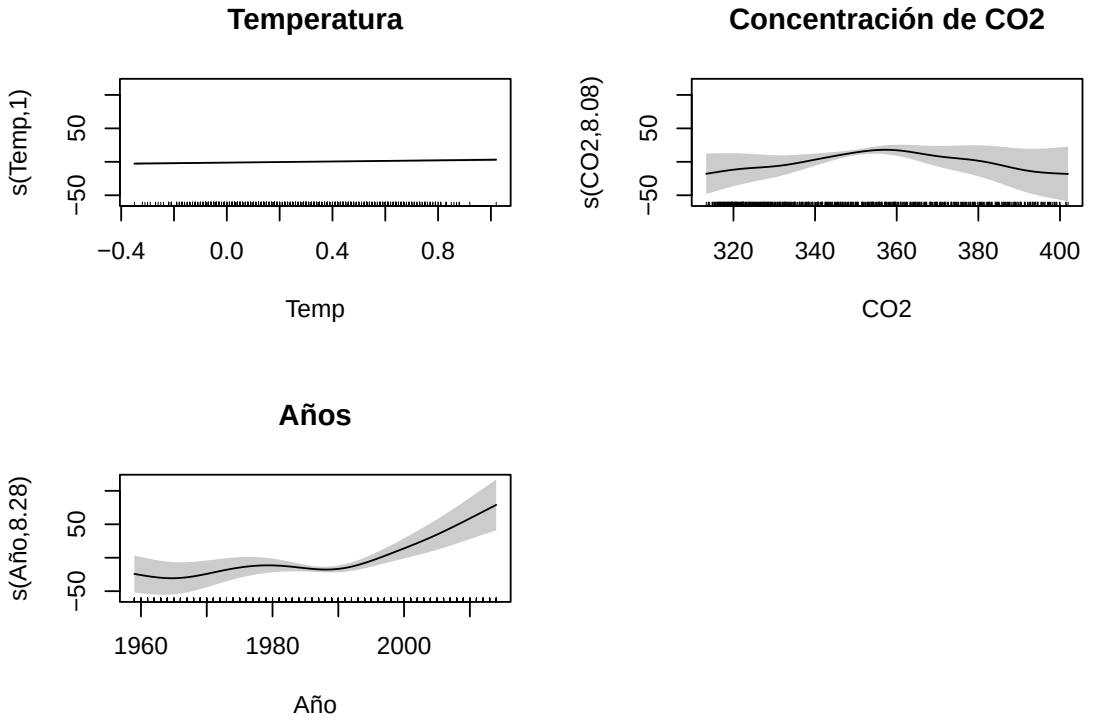
5.3.2. Descripción del modelo

Procedemos de forma similar a las secciones anteriores:

```
magSL <- gam(GMSL ~ s(Temp) + s(CO2) + s(Año) + Mes, data = Sea)
summary(magSL)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## GMSL ~ s(Temp) + s(CO2) + s(Año) + Mes
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22.9481     0.6614 -34.695  < 2e-16 ***
## Mes2         -0.6542     0.9125  -0.717 0.473643
## Mes3         -1.6037     1.0743  -1.493 0.135988
## Mes4         -2.0515     1.4304  -1.434 0.152006
## Mes5         -2.3487     1.6201  -1.450 0.147629
## Mes6         -1.7603     1.4375  -1.225 0.221208
## Mes7         -0.9714     1.0356  -0.938 0.348583
## Mes8          0.8961     0.8931   1.003 0.316067
## Mes9          2.9220     1.2168   2.401 0.016616 *
## Mes10         4.0794     1.2121   3.366 0.000809 ***
## Mes11         3.9769     0.9256   4.297    2e-05 ***
## Mes12         3.0620     0.8820   3.472 0.000552 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df      F p-value
## s(Temp) 1.000  1.000  8.967 0.00285 **
## s(CO2)  8.083  8.768 10.085 < 2e-16 ***
## s(Año)  8.279  8.820 18.497 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.98   Deviance explained = 98.1%
## GCV = 21.716  Scale est. = 20.768    n = 672
```

La desviación explicada por el modelo es del 98.1 % y el error estimado por validación cruzada generalizada es de 21.716.

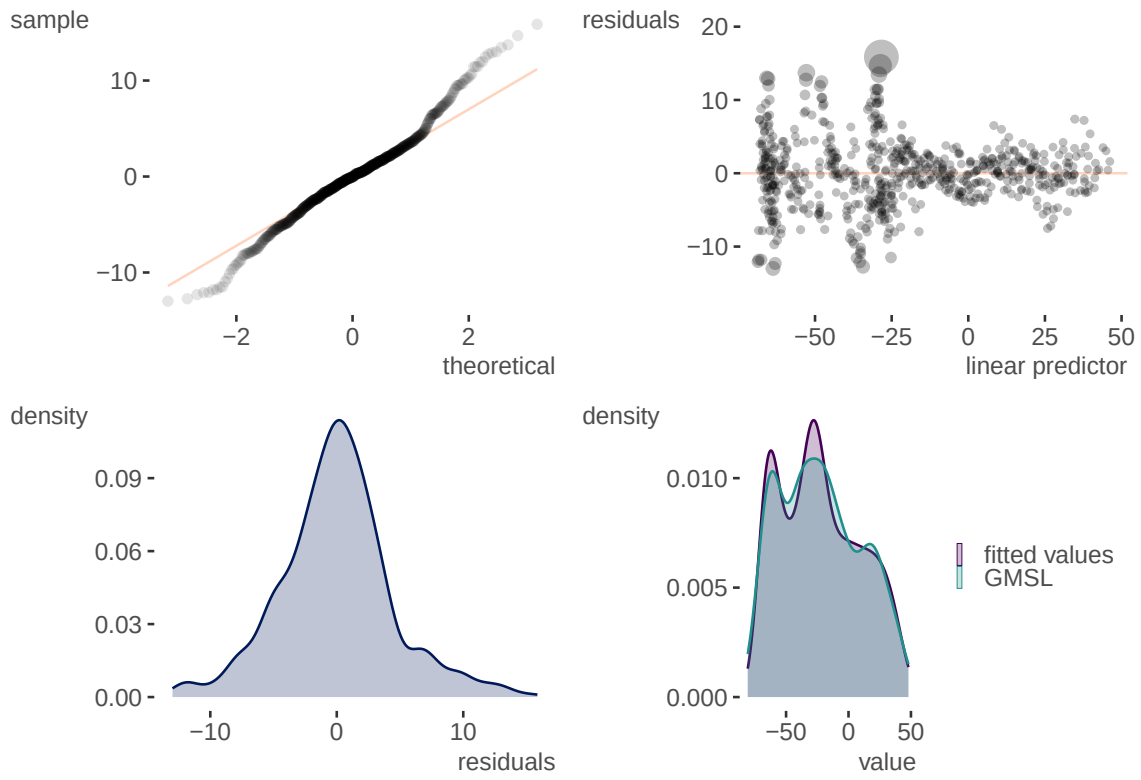


Hay indicios de que la relación de la variable *Temp* tenga un efecto lineal sobre la variable de respuesta. Lo comprobamos como hicimos en 5.1.2.

```
magSL2 <- gam(GMSL ~ Temp + s(CO2) + s(Año) + Mes, data = Sea)
anova(magSL,magSL2,test = 'F')
```

```
## Analysis of Deviance Table
##
## Model 1: GMSL ~ s(Temp) + s(CO2) + s(Año) + Mes
## Model 2: GMSL ~ Temp + s(CO2) + s(Año) + Mes
##   Resid. Df Resid. Dev          Df   Deviance      F    Pr(>F)
## 1    641.41      13346
## 2    641.41      13346 -1.7322e-05 -0.0016203 4.5043 8.293e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En este caso como el p-valor es cercano a 0 se rechaza la hipótesis nula, por lo que se trata de modelos distintos. Por comodidad trabajaremos con el primero, partimos estudiando las gráficas de diagnóstico:



En este caso nos encontramos principalmente frente a dos problemas. Por una parte, en la gráfica Q-Q plot podemos observar que en los extremos los puntos no se ajustan correctamente a la recta, se puede interpretar que se tienen colas más pesadas, es decir, los valores extremos se alejan de seguir una distribución normal. Por otra parte, en la gráfica de arriba a la derecha se puede intuir un patrón en los residuos, lo que sugiere que no se verifique la hipótesis de varianza constante.

Podemos observar qué ocurre si permitimos que haya interacción entre las variables al definir el modelo aditivo generalizado:

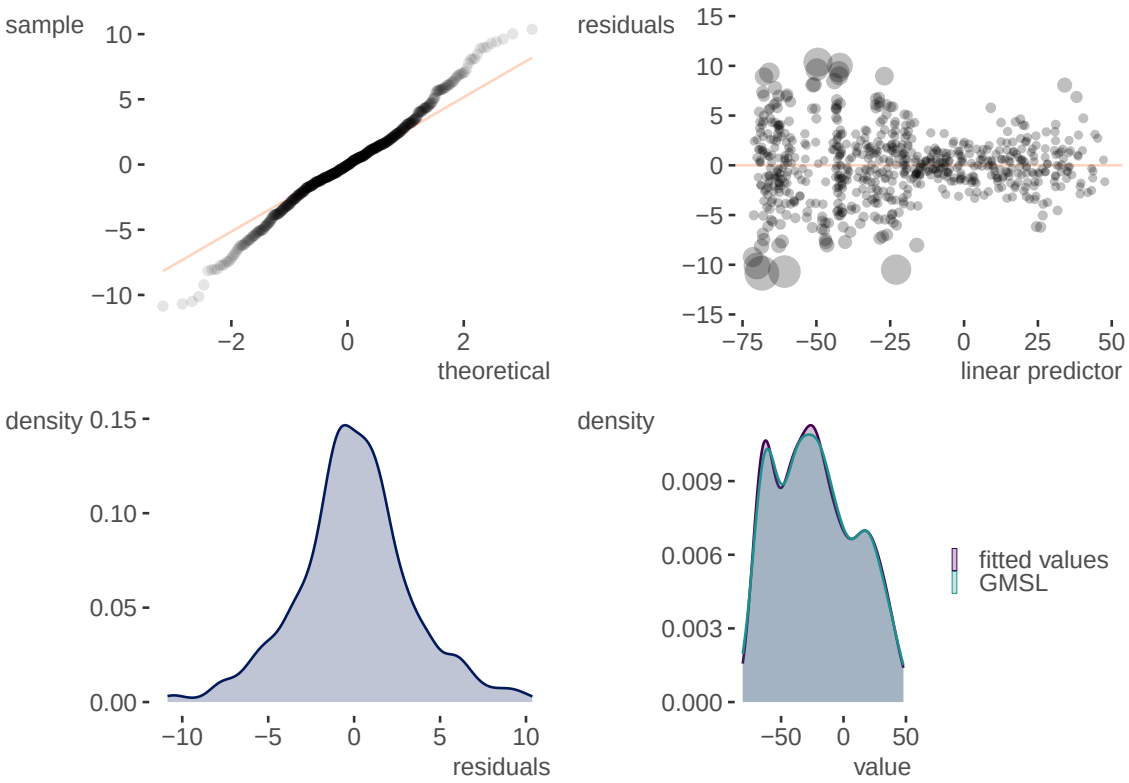
```
magSL3 <- gam(GMSL ~ s(Temp) + s(CO2) + s(Año) + Mes + s(Temp,CO2,Año),
              data = Sea)
summary(magSL3)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## GMSL ~ s(Temp) + s(CO2) + s(Año) + Mes + s(Temp, CO2, Año)
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.308e+01  6.072e-01 -38.004  < 2e-16 ***
## Mes2         5.375e-04  7.502e-01   0.001    0.999
## Mes3        -5.014e-01  9.430e-01  -0.532    0.595
```

```
## Mes4        -6.054e-01  1.392e+00  -0.435    0.664
## Mes5        -8.989e-01  1.669e+00  -0.539    0.590
## Mes6        -3.304e-01  1.397e+00  -0.237    0.813
## Mes7        -1.276e-01  8.847e-01  -0.144    0.885
## Mes8         2.806e-01  7.381e-01   0.380    0.704
## Mes9         6.551e-01  1.322e+00   0.496    0.620
## Mes10        1.916e+00  1.315e+00   1.456    0.146
## Mes11        3.330e+00  7.901e-01   4.215 2.90e-05 ***
## Mes12        3.377e+00  7.017e-01   4.812 1.91e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df      F  p-value
## s(Temp)           1.001  1.002  4.714 0.030282 *
## s(CO2)            1.000  1.000  3.109 0.078373 .
## s(Año)            1.000  1.000 11.612 0.000701 ***
## s(Temp,CO2,Año) 80.993 95.248 10.262  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.988   Deviance explained = 98.9%
## GCV = 14.843  Scale est. = 12.722    n = 672
```

A simple vista vemos que representa solo un 0.8 % mas de la desviación explicada y que el smoother para la variable CO_2 deja de ser relevante (de hecho si se retira del modelo se obtienen los mismos resultados). Observemos si se da una mejoría en los diagnósticos.

```
plot_gam_check(magSL3)
```



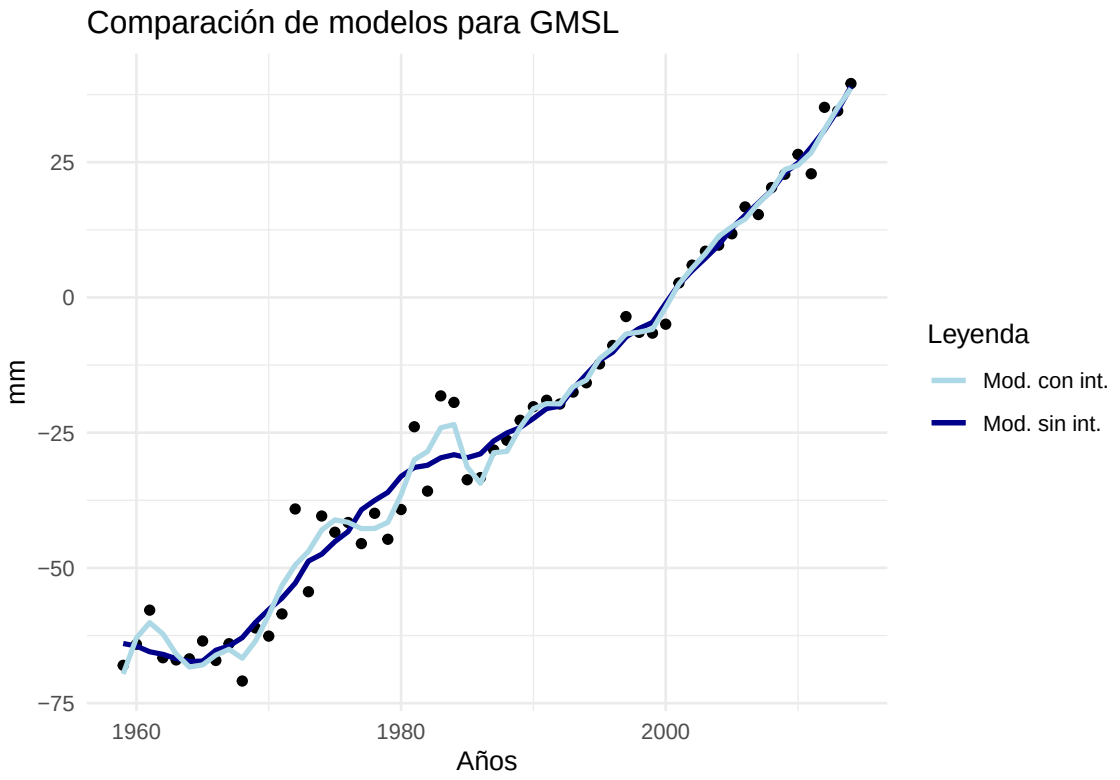
Se sigue obsevando colas pesadas en el Q-Q plot y un patrón entre los residuos y los predictores, por lo que en este aspecto el modelo no ha mejorado. Comparémoslos con los criterios AIC y GCV:

```
##                        AIC      GCV
## Sin interacción 3976.196 21.71648
## Con interacción 3706.602 14.84269
```

Bajo estos dos criterios se puede interpretar que el modelo en el que se permite la interacción entre las variables predictoras tiene un mejor desempeño que el que no lo permite.

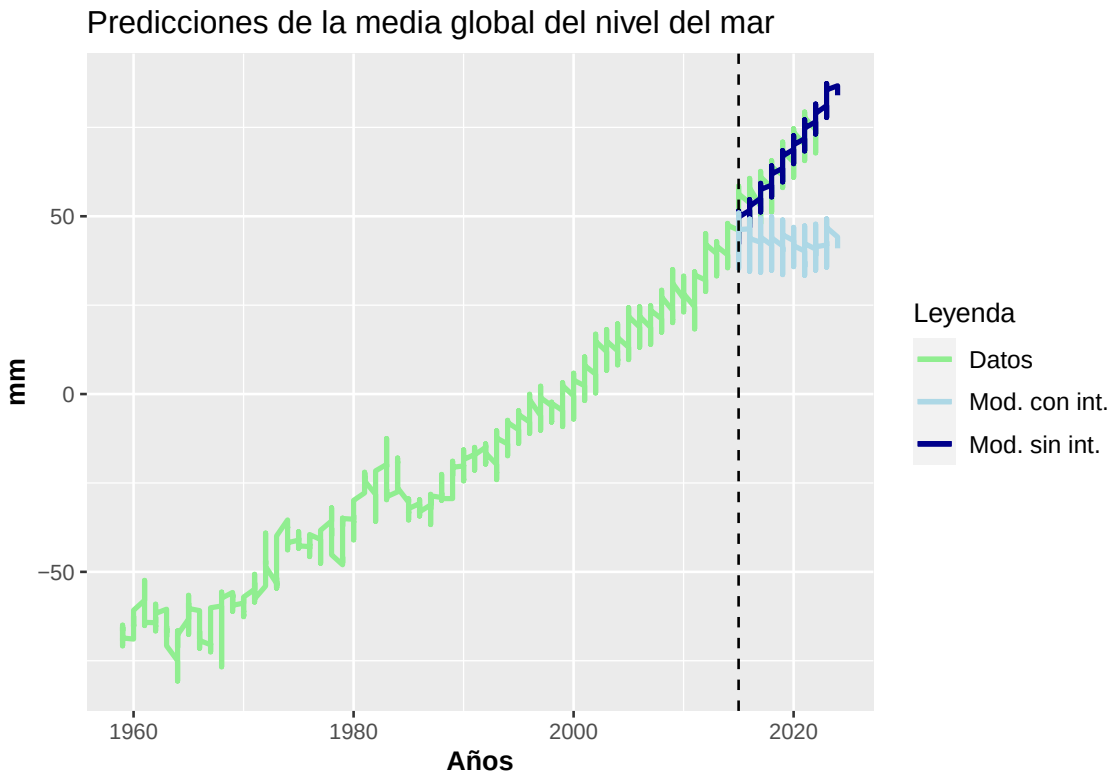
5.3.3. Visualización de resultados

Comparemos las representaciones de los datos ajustados por ambos modelos para las observaciones correspondientes a los meses de Julio desde 1959 hasta 2014. Es preferible ver cómo varía el nivel del mar para un mes fijo que representando todos los datos mensuales a la vez, ya que de tal modo se evita el ruido producido por las variaciones entre estaciones.



Se ve claramente que el modelo que admite interacción tiene más curvatura que el otro y, por ello, parece que se adapta mejor a los puntos representados. Sin embargo, ambos nos sirven para notar la subida del nivel del mar con el transcurso de los años.

Además como disponemos de los datos de *Temp* y de *CO2* hasta abril de 2024 y de *GMSL* hasta 2022, podemos también comparar cómo se adaptan los modelos a nuevas observaciones.



Se puede observar una gran diferencia entre las predicciones hechas por cada modelo. Si nos fijamos en los datos que se tenían del *GMSL* para los años desde 2015 hasta 2022, podemos ver que el modelo sin interacciones entre las variables realiza predicciones mucho más precisas que el que sí permite interacciones entre ellas. Esto puede ser debido al sobreajuste, es decir, el segundo modelo ajusta tan bien los datos disponibles que luego no es capaz de adaptarse de forma adecuada a nuevas observaciones.

Capítulo 6

Conclusión

Por un lado, durante el proyecto se ha podido observar que los modelos aditivos generalizados mejoran la metodología de los MLG incorporando la flexibilidad que aporta la regresión no paramétrica y además mantienen la interpretabilidad de los datos del análisis de regresión con múltiples variables predictoras. Actualmente, estos modelos son un punto de partida magnífico para el modelado de un problema, un MAG bien ajustado debería funcionar de manera conveniente, incluso comparado con métodos de *boosting* o de *deep learning*. Además, el MAG tiene una base para una mejor interpretabilidad y métricas de incertidumbre más sencillas, por lo que en muchos problemas asociados al análisis de datos, los Modelos Aditivos Generalizados son una buena opción.

Por otro lado, también se han presentado resultados objetivos de la existencia del cambio climático y de la gravedad de sus consecuencias, tanto en el capítulo 4 con el apoyo de artículos e invetigaciones realizadas por organizaciones, como en el capítulo 5 con tres aplicaciones de los modelos aditivos generalizados. La primera y la tercera nos sirven para comprobar algunos de los impactos que el cambio climático está teniendo sobre el planeta y la segunda nos permite analizar la tendencia de los gases de efecto invernadero en las últimas décadas. Con los datos que se manejan no es posible asegurar el efecto que la humanidad sobre este tema, sin embargo, nos podemos referir a los informes del IPCC (Climático) para obtener resultados objetivos que aseguran las implicaciones que tiene la actividad humana en él. Además, en estos informes podemos encontrar estudios de los posibles escenarios y estrategias para la mitigación o la adaptación a ellos.

En conclusión, el uso de modelos aditivos generalizados no solo ayuda a la comprensión del cambio climático, sino que también proporciona una base sólida para la toma de decisiones y la planificación a largo plazo de métodos para su desaceleración. La flexibilidad y precisión que ofrecen los MAG los convierten en una herramienta invaluable para abordar los desafíos impuestos por uno de los problemas más urgentes de nuestro tiempo, el cambio climático.

Apéndice A

Apéndice: Carga y depuración de datos

Clima arepuerto SVQ

5.1.1

```
library(tidyr)
library(dplyr)
Clima <- aemet_monthly_period(station = "5783", start = 1960, end = 2023)
Clima <- Clima %>% separate(fecha, into = c("Año", "Mes"), sep = "-")
Clima$Año <- as.numeric(Clima$Año)
Clima$Mes <- factor(Clima$Mes, levels = as.character(1:12))
Clima <- Clima[,c(1,2,6,11,27,29,32)] # Seleccionamos las variables que
# nos interesan
colnames(Clima) <- c('Año','Mes','HR','PresM','Prec','WMed','TMedM')
Clima <- Clima %>% arrange(Año, Mes) # Ordenamos por año y mes
Clima <- Clima[complete.cases(Clima$Mes),] # Retiramos las medias anuales
```

Concentración atmosférica de CO2

5.2.1

```
CO2 <- read_excel('trends-in-atmospheric-carbon-dioxide-concentration.xlsx')

CO2$DateTime <- as.Date(CO2$DateTime)
CO2$Año <- as.numeric(year(CO2$DateTime))
CO2$Mes <- factor(month(CO2$DateTime),levels = as.character(1:12))
CO2$Tmes <- as.numeric(CO2$'Monthly Data')
CO2$Trend <- as.numeric(CO2$'Trend')
CO2 <- CO2[,c(4,5,6,3)]
CO2 <- CO2 %>% arrange(Año, Mes)
```

Concentración atmosférica de CH4

5.2.1

```
CH4 <- read_excel('trends-in-atmospheric-methane-concentration.xlsx')

CH4$DateTime <- as.Date(CH4$DateTime)
CH4$Año <- as.numeric(year(CH4$DateTime))
CH4$Mes <- factor(month(CH4$DateTime),levels = as.character(1:12))
CH4$Trend <- as.numeric(CH4$'Trend')
CH4 <- CH4[,c(3,4,2)]
CH4 <- CH4 %>% arrange(Año, Mes)
```

Concentración atmosférica de N2O

5.2.1

```
N2O <- read_excel('trends-in-atmospheric-nitrous-oxide-concentration.xlsx')

N2O$DateTime <- as.Date(N2O$DateTime)
N2O$Año <- as.numeric(year(N2O$DateTime))
N2O$Mes <- factor(month(N2O$DateTime),levels = as.character(1:12))
N2O$Trend <- as.numeric(N2O$'Trend')
N2O <- N2O[,c(3,4,2)]
N2O <- N2O %>% arrange(Año, Mes)
```

Global Mean Sea Level

5.3.1

```
SeaL <- read_excel('SeaLevelv2.xlsx')

# Retiramos datos con otro formato y nos quedamos con las variables
# necesarias.
SeaL <- SeaL[-(1:240),c(8,9,11)]
# La variable Año es numérica para la representación como ya se
# ha indicado
# otras veces
# y la mensual es categórica
SeaL$Year <- as.numeric(SeaL$Year)
SeaL$Month <- factor(SeaL$Month,levels = 1:12)
colnames(SeaL) <- c('Año','Mes','GMSL')

# Como en algunos casos se tienen varias mediciones por mes lo
# que hacemos es tomar como la GMSL mensual la media de todas ellas.
SeaL <- SeaL %>%
  group_by(Año, Mes) %>%
  summarise(GMSL = mean(GMSL, na.rm = TRUE))
summary(SeaL)
```

Temperatura media global al nivel del mar

5.3.1

```
library(tidyverse)

raw_data <- read_excel("GSST.xlsx", col_names = FALSE)

# Convertimos el data frame a un vector de caracteres
data_vector <- as.vector(raw_data[[1]])

# Separamos los valores por comas
processed_data <- strsplit(data_vector, split = ",")

# Convertimos la lista resultante en un data frame
data_frame <- do.call(rbind, lapply(processed_data,
                                    function(x) as.numeric(x)))

colnames(data_frame) <- c("Año",
                          paste0("Temp", 1:(ncol(data_frame) - 1)))

# Convertimos a tibble para facilitar la manipulación
data_tibble <- as_tibble(data_frame)

# Transponemos data_tibble a formato vertical
data_long <- data_tibble %>%
  pivot_longer(cols = starts_with("Temp"), names_to = "Mes",
               values_to = "Temp") %>%
  mutate(Mes = as.numeric(gsub("Temp", "", Mes))) %>%
  arrange(Año, Mes)

SeaT <- data_long[!(data_long$Mes >12),]
summary(SeaT)
```

Unión de los conjuntos de datos para el modelo del nivel del mar

```
SeaLb <- (SeaL[(2015 > SeaL$Año) &(SeaL$Año >1958),])[1:672,]
SeaTb <- (SeaT[(2015 > SeaT$Año) & (SeaT$Año > 1958),])[1:672,]
CO2b <-  CO2[(2015 > CO2$Año) & (CO2$Año > 1958),]

Sea <- cbind(SeaLb,SeaTb,CO2b)
Sea <- Sea[,c(1,2,3,6,9)]
colnames(Sea) <- c('Año','Mes','GMSL','Temp','CO2')
str(Sea)
summary(Sea)
```

Apéndice B

Apéndice: Representaciones gráficas

Representaciones de la primera aplicación 5.1.2

```
par(mfrow = c(2, 2))
plot(mag1, select = 1, main = 'Humedad Relativa', shade = TRUE)
plot(mag1, select = 2, main = 'Presion', shade = TRUE)
plot(mag1, select = 3, main = 'Precipitaciones totales', shade = TRUE)
plot(mag1, select = 4, main = 'Velocidad media del viento', shade = TRUE)
par(mfrow = c(1, 1))
```

```
Julio <- filter(Clima, Clima$Mes == 7)
Julio$Preds <- predict(mag1, newdata = Julio)
Julio <- Julio[complete.cases(Julio$Preds),]

lm1 <- gam(TMedM ~ Año,data = Julio)
Julio$LPreds <- predict(lm1,Julio)

library(ggplot2)
ggplot(Julio,aes(x=Año))+
  geom_point(aes(y=TMedM),size=1.5, col = 'black')+
  theme_minimal()+
  geom_line(aes(y=Preds, color = 'MAG'),linewidth=1)+
  geom_line(aes(y=LPreds, color = 'ML'),linewidth=0.6)+
  labs(title ="Temperatura media mensual del mes de Julio",x="Año",
       y="Temperatura ºC")+
  scale_color_manual(values = c('MAG' = 'darkred', 'ML' = '#EB6146'),
                     name = "Leyenda") +
  theme(axis.title = element_text(face = "bold"),
        legend.text = element_text( size = 10,colour = "black"),
        legend.position = 'right')
```

Representaciones de la segunda aplicación 5.2.2

```r
predsc <- exp(predict(magCO2c,CO2))

lmCO2 <- gam(Tmes ~ Año, )

ggplot(CO2, aes(x = Año, y = Tmes)) +
  geom_point(size = 2) +
  geom_line(aes(y = predsc), color = "darkgreen", linewidth = 1.2) +
  labs(x = "Años",y = "ppm",
       title = "Concentración mensual media de dióxido de carbono (ppm)") +
  theme_minimal()
```

```r
futuro = CO2[1:324,1:2]
s <- c()
new <- function(x){
  for (i in x){
    for (j in 1:12){
      s <- append(s,i)
    }
  }
s}
futuro$Año <- new(2024:2050)

futuro$Tmes <- sqrt(1/predict(magCO2b,futuro))
futuro$Trend <- 1:324 == NA
futuro$col <- 'Predicciones'

pasado <- CO2
pasado$col <- 'Datos'
df <- rbind(pasado,futuro)

ggplot(df, aes(x = Año, y = Tmes)) +
  geom_line(aes(color = col, group = 1), linewidth = 1) +
  geom_vline(aes(xintercept = 2024), color = 'black',
             linetype = "dashed" , linewidth = 0.5)+
  labs(x = "Años",y = "ppm",
       title = "Concentración mensual media de dióxido de carbono",
       legend = c('Datos','Predicciones')) +
  scale_color_manual(values = c('Datos' = 'darkgreen',
                                'Predicciones' = 'darkred'),
                     name = "Leyenda")+
  theme(axis.title = element_text(face = "bold"),
        legend.text = element_text(size = 10,color = 'black'),
        legend.position = 'right')
```

```r
magCH4 <- gam(Trend ~ s(Año) + as.numeric(Mes),
              data = CH4, family = Gamma(link = "log"))
```

```r
magN2O <- gam(Trend ~ s(Año) + Mes,
              data = N2O, family = Gamma(link = "log"))


predsCH4 <- exp(predict(magCH4,CH4))
predsN2O <- exp(predict(magN2O,N2O))

par(mfrow = c(1,2))

ggplot(CH4, aes(x = Año, y = Trend)) +
  geom_point(size = 2) +
  geom_line(aes(y = predsCH4), color = "darkorange2", linewidth = 1.2) +
  labs(x = "Años",y = "ppb", title = "Concentración mensual media
       de metano") +
  theme_minimal()

ggplot(N2O, aes(x = Año, y = Trend)) +
  geom_point(size = 2) +
  geom_line(aes(y = predsN2O), color = "deepskyblue2", linewidth = 1.2) +
  labs(x = "Años",y = "ppb", title = "Concentración mensual media
       de óxido nitroso") +
  theme_minimal()

par(mfrow = c(1,1))
```

Representaciones de la tercera aplicación 5.3.2

```r
par(mfrow = c(2, 2))
plot(magSL, select = 1, main = 'Temperatura', shade = TRUE)
plot(magSL, select = 2, main = 'Concentración de CO2', shade = TRUE)
plot(magSL, select = 3, main = 'Años', shade = TRUE)
par(mfrow = c(1, 1))
```

```r
Julio <- Sea[Sea$Mes == 7,]
Julio$preds <- predict(magSL, newdata = Julio)
Julio$preds3 <- predict(magSL3, newdata = Julio)

ggplot(data = Julio, aes(x = Año, y = GMSL)) +
  geom_point() +
  geom_line(aes(x = Año, y = preds, color = 'Mod. sin int.'),
            linewidth = 1) +
  geom_line(aes(x = Año, y = preds3, color = 'Mod. con int.'),
            linewidth = 1) +
  labs(x = "Años", y = "mm",
       title = "Comparación de modelos para GMSL") +
  scale_color_manual(values = c('Mod. sin int.'= "darkblue",
                                'Mod. con int.' = "lightblue"),
```

```
                    name = 'Leyenda') +
  theme_minimal()
```

```
SeaTpreds <- (SeaT[(2015 <= SeaT$Año),])[1:112,]
CO2preds <-  CO2[(2015 <= CO2$Año),]

Seapreds <- cbind(SeaTpreds,CO2preds)
Seapreds <- Seapreds[,c(1,2,3,6)]
colnames(Seapreds) <- c('Año','Mes','Temp','CO2')

na_rows <- as.data.frame(matrix(NA, nrow = 112, ncol = ncol(Sea)))
colnames(na_rows) <- colnames(Sea)

SeaP <- rbind(Sea, na_rows)

SeaP$preds <- append(1:672 == NA,predict(magSL,Seapreds))
SeaP$preds3 <- append(1:672 == NA,predict(magSL3,Seapreds))
SeaP$Año <- append(Sea$Año,Seapreds$Año)
SeaP$Mes <- append(Sea$Mes,Seapreds$Mes)
SeaP$GMSL <- append(SeaL$GMSL[(1959 <= SeaL$Año)],rep(NA, 16))

ggplot(data = SeaP, aes(x = Año, y = GMSL)) +
  geom_line(aes(x = Año, y = GMSL , color = 'Datos'),
            linewidth = 1) +
  geom_line(aes(x = Año, y = preds, color = 'Mod. sin int.'),
            linewidth = 1) +
  geom_line(aes(x = Año, y = preds3, color = 'Mod. con int.'),
            linewidth = 1) +
  labs(x = "Años", y = "mm",
       title = "Predicciones de la media global del nivel del mar") +
  geom_vline(aes(xintercept = 2015), color = 'black',
             linetype = "dashed" , linewidth = 0.5)+
  scale_color_manual(values = c('Datos' = 'lightgreen',
                                'Mod. sin int.' = 'darkblue',
                                'Mod. con int.' = 'lightblue'),
                     name = "Leyenda")+
  theme(axis.title = element_text(face = "bold"),
        legend.text = element_text(size = 10,color = 'black'),
        legend.position = 'right')
```

Bibliografía

Deny, deceive, delay (vol 3): Climate information integrity ahead of cop28. nov 2023. URL https://caad.info/wp-content/uploads/2023/11/Deny-Deceive-Delay-Vol.-3-1.pdf.

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

Sophie Boehm and Clea Schumer. 10 big findings from the 2023 ipcc report on climate change. mar 2023. URL https://www.wri.org/insights/2023-ipcc-ar6-synthesis-report-climate-change-findings#.

Nunez Christina. ¿qué son los gases de efecto invernadero y cuáles son sus efectos? URL https://www.nationalgeographic.es/medio-ambiente/gases-efecto-invernadero-que-son-hacen.

Michale Clark. Generalized additive models.

Oficina Española Cambio Climático. Informes de evaluación del ipcc. URL https://www.miteco.gob.es/es/ceneam/recursos/mini-portales-tematicos/cclimatico/informe_ipcc.html.

Caroline Craig. Sea level rise 101. 2024. URL https://www.nrdc.org/stories/sea-level-rise-101#what-is.

Pedro Delicado. Curso de modelos no paramétricos. URL https://www.researchgate.net/publication/267795562_Curso_de_Modelos_no_Parametricos.

Unión Europea. Comisión europea - cambio climático. URL https://climate.ec.europa.eu/climate-change_es.

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

IPCC. Climate change 2022: Impacts, adaptation, and vulnerability., 2022. URL https://www.ipcc.ch/report/ar6/wg2/chapter/summary-for-policymakers/.

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

Lottie Limb. From oil ads to russian interference: These are the climate lies to look out for at cop28. nov 2023. URL https://www.euronews.com/green/2023/11/29/from-russian-pr-to-online-grifters-who-is-peddling-climate-disinformation-in-2023.

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

Krista F Myers, Peter T Doran, John Cook, John E Kotcher, and Teresa A Myers. Consensus revisited: quantifying scientific agreement on climate change and climate expertise among earth scientists 10 years later. *Environmental Research Letters*, 16 (10):104030, oct 2021. doi: 10.1088/1748-9326/ac2774. URL https://dx.doi.org/10.1088/1748-9326/ac2774.

Manuel Pizarro, Diego Hernangómez, and Gema Fernández-Avilés. *climaemet: Climate AEMET Tools*, 8 2021. URL https://hdl.handle.net/10261/250390.

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

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

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

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

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