

Fundamentos de ciencia de datos con R

Gema Fernández-Avilés y José-María Montero

2023-06-01

Índice general

Prefacio	5
¡Hola mundo!	5
¿Por qué este libro?	6
¿A quién va dirigido?	7
El paquete CDR	8
¿Por qué R?	8
Agradecimientos	9
1. Modelos aditivos generalizados	11
1.1. Introducción	11
1.2. Splines con penalizaciones	12
1.3. Aspectos metodológicos	13
1.4. Procedimiento con R: la función <code>gam()</code> del paquete <code>mgcv</code>	16
1.5. Casos prácticos	17

Prefacio

¡Hola mundo!

El siglo XXI está siendo testigo de grandes cambios vertiginosos en el contexto social y tecnológico, entre otros. Los tiempos han cambiado, la sociedad se ha globalizado y “exige” respuestas inmediatas a problemas muy complejos. Vivimos en el mundo de la **información**, de los **datos**, o mejor, de las **bases de datos masivas**, y los ciudadanos y, sobre todo, las empresas y los gobiernos, dirigen su mirada hacia el mundo científico para que les ayude a “**oír las historias**” que cuentan esos datos acerca de la realidad de la que han sido extraídos. Y dado su enorme volumen y sofisticación (en el nuevo mundo las imágenes y los textos, por ejemplo, también son datos), exigen algoritmos de nueva generación en el campo del *machine learning*, o incluso del *deep learning*, para “oír las historias” que cuentan. No parecen mirar al “antiguo” investigador científico, sino al “nuevo” *científico de datos*.

Ello, inevitablemente, se traduce en la necesidad de profesionales con una gran capacidad de adaptación a este nuevo paradigma: los científicos de datos, también llamados por algunos los “nuevos hombres del Renacimiento”, para lo cual las Universidades y demás instituciones educativas especializada se apresuran a incluir el grado de Ciencia de Datos en su oferta educativa y a ofrecer seminarios de software estadístico de acceso abierto para sus estudiantes de primeros cursos.

Con la emergencia de la nueva sociedad, en la que el manejo de la ingente cantidad de información que genera se hace absolutamente necesario para circular por ella, la **Ciencia de Datos** ha venido para quedarse. Sin embargo, el mundo de la Ciencia de Datos es cualquier cosa menos sencillo. En él, cualquier ayuda, cualquier guía es bienvenida. Por ello, es muy recomendable que la persona que se quiera introducir en él, sea con fines de investigación o con fines profesionales, se agarre de la mano de un guía especializado que le lleve, de una manera amena, comprensible y eficiente, desde el planteamiento de su problema y la captura de la información necesaria para poderle dar una solución, hasta la redacción de las conclusiones finales que ha obtenido con los modernos informes reproducibles colaborativos. Y como en la parte central de ese camino tendrá que luchar con grandes gigantes (en la actualidad denominados técnicas estadísticas y algoritmos), el guía tendrá que explicarle, de manera sencilla y amena, en qué consiste la lucha (las técnicas y los algoritmos) y cómo llegar a la victoria lo más rápido posible, enseñándole a moverse por el mundo del software estadístico, en nuestro caso **R**, que le permitirá realizar los cálculos necesarios para vencer al problema planteado a una velocidad vertiginosa.

En resumen, la información masiva y el moderno tratamiento estadístico de la misma son la “mano invisible” que gobierna la sociedad del siglo XXI, y este manual pretende ser el guía anteriormente mencionado que le llevará de la mano cuando quiera caminar por ella.

¿Por qué este libro?

Lo dicho anteriormente ya justifica por sí solo la aparición de este manual. Afortunadamente, no es el primero en la materia, pues son ya bastantes los materiales de calidad publicados sobre Ciencia de Datos. Sin embargo, quizás, éste pueda ser considerado el más completo. Y ello por varias razones.

La primera es su **completitud**: este manual lleva de la mano al lector desde el planteamiento del problema hasta el informe que contiene la solución al mismo; o desde no saber qué hacer con la información de la que dispone, hasta ser capaz de transformar tales bases de datos masivas, y casi imposibles de manejar, en respuestas a problemas fundamentales de una empresa, institución o cualquier agente social.

La segunda es su **amplitud temática**:

- (I) Parte de las dos primeras preguntas que un neófito se puede hacer sobre esta temática: ¿qué es eso de la Ciencia de Datos que está en boca de todos? Y, ¿qué diablos es **R** y cómo funciona?
- (II) Enseña cómo moverse en la jungla del *Big Data* y de los “nuevos” tipos de datos, siempre bajo el paraguas de la ética de los datos y del buen gobierno de dichos datos.
- (III) Muestra al lector cómo obtener conocimiento de la oscuridad del enorme banco de información a su disposición, que no sabe cómo abordar ni manejar.
- (IV) No deja a nadie atrás, y de forma previa al contenido central del manual (las técnicas de Ciencia de Datos), incluye unas breves, pero magníficas, secciones sobre los rudimentos de la probabilidad, la inferencia estadística y el muestreo, para aquéllos no familiarizados con estas cuestiones.
- (V) Aborda una treintena de técnicas de Ciencia de Datos en el ámbito de la modelización, análisis de datos cualitativos, discriminación, *machine learning* supervisado y no supervisado, con especial incidencia en las tareas de clasificación y clusterización -así como, en el caso no supervisado, de reducción de la dimensionalidad, escalamiento multidimensional y análisis de correspondencias-, *deep learning*, análisis de datos textuales y de redes, y, finalmente, ciencia de datos espaciales (desde las perspectivas de la geoestadística, la econometría espacial y los procesos de punto).
- (VI) Hace especial hincapié en la reproducibilidad en tiempo real (o no) entre los distintos miembros de un equipo (sea universitario, empresarial, o del tipo que sea) y en la difusión de los resultados obtenidos, enseñando al lector cómo generar informes reproducibles mediante RMarkdown y documentos Quarto o en otros modernos formatos.
- (VII) Dedica un capítulo a la creación de aplicaciones web interactivas (con Shiny).

Índice general

7

- (viii) Para aquéllos con pasión por la codificación, y que quieran compartir código y colaborar con otros desarrolladores, este manual aborda la gestión rápida y eficaz de proyectos (del tamaño que sean) mediante Git, un sistema de control de versiones distribuido, gratuito y de código abierto, y GitHub, un servicio de alojamiento de repositorios Git del cual, aquellos no familiarizados con la cuestión de la codificación, o con aversión a ella, podrán tomar el código que necesitan.
- (ix) Muestra al lector los primeros pasos para iniciarse en el geoprocесamiento en la nube.
- (x) Y, finalmente, aborda más de una docena de casos de uso (en medicina, periodismo, economía, criminología, marketing, moda, demanda de electricidad, cambio climático, reconocimiento de patrones en la forma de tuitear...) que ilustran la puesta en práctica de todos los conocimientos anteriormente adquiridos.

La cuarta razón es que todo lo que el lector aprende en este manual lo puede reproducir y poner en práctica inmediatamente con **R**, puesto que el manual está trufado de *chunks* (o trozos de código **R**) que no tiene más que cortar y pegar para reproducir los ejemplos que se muestran en el libro, cuyos datos están en el paquete CDR; o utilizar dichas *chunks* para abordar el problema que le ocupa con los datos que tenga a su disposición. Una buena razón, sin duda. Por consiguiente, el manual es una buena combinación “teoría-práctica-software” que permite abordar cualquier problema que el científico de datos se plante en cualquier disciplina o situación empresarial, médica, periodística...

La quinta es su **variedad de perspectivas**. Son **más de 40 los participantes** en este manual. Algunos de ellos, prestigiosos profesores universitarios; otros, destacados miembros de instituciones públicas; otros, CEOs de empresas en la órbita de la ciencia de datos; otros, *big names* del mundo de **R** software... El manual es, sin duda, un magnífico ejemplo de colaboración Universidad-Empresa para buscar soluciones a los problemas de las sociedades modernas.

¿A quién va dirigido?

Fundamentos de ciencia de datos con R está dirigido a todos aquellos que desean desarrollar las habilidades necesarias para abordar proyectos complejos de Ciencia de Datos y “pensar con datos” (como lo acuñó Diane Lambert, de Google). El deseo de resolver problemas utilizando datos es su piedra angular. Por tanto, como se avanzó anteriormente, este manual no deja a nadie atrás, y lo único que requiere es “el deseo de resolver problemas utilizando datos”. No excluye ninguna disciplina, no excluye a las personas que no tengan un elevado nivel de análisis estadístico de datos, no excluye a nadie. Se ha procurado una combinación de rigor y sencillez, y de teoría y práctica, todo ello con sus correspondientes códigos en **R**, que satisfaga tanto a los más exigentes como a los principiantes.

También está destinado a todos aquellos que quieran sustituir la navegación por la web (la búsqueda del video, publicación de blog o tutorial *online* que solucione su problema –frustración tras frustración por la falta de consistencia, rigor e integridad de dichos materiales, así como por su sesgo hacia paquetes singulares para la implementación de las cuestiones que tratan–), por

una “**biblia de la ciencia de datos**” rigurosa pero sencilla, práctica y de aplicación inmediata sin ser ni un experto estadístico ni un experto informático.

Pero si a alguien está destinado especialmente, es a la comunidad hispano hablante. Este manual es un guiño a dicha comunidad, para que tenga a su disposición, en su lengua nativa, uno de los mejores manuales de Ciencia de Datos de la actualidad.

El paquete CDR



El paquete **CDR** contiene la mayoría de conjuntos de datos utilizados en este libro que no están disponibles en otros paquetes. Para instalarlo use la función `install_github()` del paquete `remotes`.

```
# este comando sólo necesita ser ejecutado una vez
# si el paquete remotes no está instalado, descomentar para instalarlo

# install.packages("remotes")
remotes::install_github("cdr-book/CDR")
```

La lista de todos los conjuntos de datos puede obtenerse haciendo `data()`.

```
library('CDR')
data(package = "CDR")
```

Este paquete ayudará al lector a reproducir todos los ejemplos del libro. De acuerdo con las mejores prácticas en **R**, el paquete **CDR** sólo contiene los datos utilizados en el libro.

¿Por qué R?

R es un lenguaje de código abierto para computación estadística que se ha consolidado entre la comunidad científica internacional, en las últimas dos décadas, como una herramienta de primer

Índice general

9

nivel, consolidándose como líder permanente en el ámbito de la implementación de metodologías estadísticas para el análisis de datos. La utilidad de **R** para la Ciencia de Datos deriva de un fantástico ecosistema de paquetes (activo y en crecimiento), así como de un buen elenco de otros excelentes recursos: libros, manuales, *blogs*, foros y *chats* interactivos en las redes sociales, y una gran comunidad dispuesta a colaborar, a orientar y a resolver diferentes cuestiones relacionadas con **R**.

Por otra parte, **R** es el lenguaje estadístico y de análisis de datos más utilizado en la mayoría de los entornos académicos y, cómo no, por una larga lista de importantes empresas, entre las que se cuentan Facebook (análisis de patrones de comportamientos relacionado con actualizaciones de estado e imágenes de perfil), Google (para la efectividad de la publicidad y la previsión económica), Twitter (visualización de datos y agrupación semántica), Microsoft (adquirió la empresa Revolution R), Uber (análisis estadístico), Airbnb (ciencia de datos), IBM (se unió al grupo del consorcio R), New York Times (visualización)...

La comunidad **R** también es particularmente generosa e inclusiva, y hay grupos increíbles, como *R-Ladies* y *Minority R Users*, diseñados para ayudar a garantizar que todos aprendan y usen las capacidades de **R**.

Agradecimientos

No queremos dar por finalizado este prefacio sin agradecer a los 44 autores participantes en esta obra su esfuerzo por condensar, en no más de 20 páginas, la teoría, práctica y tratamiento informático de la parte de la Ciencia de Datos que les fue encargada. Y no sólo eso; el “más difícil todavía” fue que debían dirigirse a un abanico de potenciales lectores tan grande como personas haya con “el deseo de resolver problemas utilizando datos”. Era misión imposible. Sin embargo, a la vista del resultado, ha sido misión cumplida. El esfuerzo mereció la pena.

Además, nos gustaría agradecer el apoyo incondicional recibido por (en orden alfabético): Itzcoatl Bueno, Ismael Caballero, Emilio L. Cano, Diego Henangómez, Ricardo Pérez, Manuel Vargas y Jorge Velasco.

También queremos poner de manifiesto que la edición de este texto ha sido financiada por diversos entes de la Universidad de Castilla-La Mancha. En su mayor parte, por el **Máster en Data Science y Business Analytics (con R software)** (a través de la orgánica: 02040M0280), pero también por la Facultad de Ciencias Jurídicas y Sociales de Toledo (a través de su contrato programa: orgánica 00440710), el Departamento de Economía Aplicada I (mediante sus fondos departamentales, DEAI 00421I126) y el Grupo de Investigación Economía Aplicada y Métodos Cuantitativos (que ha dedicado parte de sus fondos a la edición de esta obra, orgánica 01110G3044-2023-GRIN-34336).

A todos, eternamente agradecidos por ayudarnos en este reto de transformar la oscuridad en conocimiento, de convertir en una ciencia y en un arte la difícil tarea de sacar valor de los datos, el petróleo del futuro. Quizás en este momento no seamos conscientes de que hemos puesto nuestro granito de arena a la ciencia que, a buen seguro, juegue uno de los papeles más importantes de este siglo, caracterizado por el predominio de la información. Una ciencia, la Ciencia de Datos, que combina el análisis estadístico de datos, la algoritmia y el conocimiento del

negocio para sacar valor del bien más abundante de la sociedad en la que vivimos: la información. Una disciplina cuyo dominio caracteriza a los científicos de datos (también denominados los nuevos personajes del Renacimiento), profesión que ya fue calificada hace más de veinte años en la *Harvard Business Review* y en *The New York Times*, entre otros, como la “más sexy del siglo XXI”.

Nota

Este manual está publicado por [McGraw Hill](#). Las copias físicas están disponibles en [McGraw Hill](#). La versión *online* se puede leer de forma gratuita en <https://cdr-book.github.io/> y tiene la [licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional](#).

Si tiene algún comentario o sugerencia, no dude en contactar con los editores y los autores. ¡Gracias!

Capítulo 1

Modelos aditivos generalizados

María Durbán^a y Víctor Casero-Alonso^b

^aUniversidad Carlos III de Madrid ^bUniversidad de Castilla-La Mancha

1.1. Introducción

Los modelos lineales, o los lineales generalizados (GLM), vistos en los capítulos ?? y ?? tienen la ventaja de ser fáciles de ajustar e interpretar. Además, se dispone de técnicas para contrastar las hipótesis del modelo. Sin embargo, cuando la variable respuesta no está relacionada de forma lineal con las variables explicativas no tiene sentido utilizar modelos lineales (generalizados o no) y hay que acudir a modelos que flexibilicen esta relación, que, en el caso de una única variable explicativa, se puede expresar como sigue:

$$Y = \beta_0 + f(X) + \varepsilon.$$

Puede que la función $f()$ sea conocida de antemano, como ocurre en muchos modelos biológicos, donde existe una dependencia de tipo exponencial, $f(x) = e^{\beta_0 + \beta_1 x}$. En otras ocasiones, dicha función es desconocida y se puede utilizar una aproximación. Por ejemplo, mediante la regresión polinómica, muy utilizada en la práctica:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_p X^p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2). \quad (1.1)$$

Sin embargo, la regresión polinómica tiene un gran inconveniente: que no se lleva a cabo de forma local y, por tanto, cada vez que se cambia un coeficiente del modelo, el cambio impacta a los valores ajustados en todo el rango de la variable explicativa. No obstante, es posible utilizar técnicas (como las que se presentan en este capítulo) en las que el valor predicho en un punto dado sólo depende de las observaciones en ese punto y de las observaciones vecinas, es decir, el ajuste se lleva a cabo de forma local.

En el caso de disponer de más de una variable explicativa, la extensión del modelo de regresión múltiple sería el **modelo aditivo** (en el caso de variable respuesta gaussiana), donde no se asume que la relación entre la variable respuesta y cada una de las variables explicativas tenga que ser lineal:

$$Y = f(X_1) + \dots + f(X_j) + \epsilon, \quad \epsilon \sim N(0, \sigma^2). \quad (1.2)$$

Las funciones $f()$ incluyen también a las funciones lineales vistas en el Capítulo ???. Los **modelos aditivos generalizados**, GAM, extienden el modelo anterior a respuestas no gaussianas, como lo hacen los GLM respecto de los modelos lineales con respuesta gaussiana (véase Cap. ??).

1.2. Splines con penalizaciones

Las funciones de la Eq. (1.2) se estiman mediante técnicas de suavizado o *smoothers*, cuyo objetivo es extraer las tendencias (o señales) existentes en la relación entre la variable respuesta y las variables explicativas, sin presuponer una forma funcional a priori para ellas; sólo se asume que la relación entre Y y X es suave (tiene poco ruido). Las predicciones obtenidas mediante estas técnicas tienen menos variabilidad que la variable respuesta; de ahí que a estas técnicas se les denomine “suavizadores” (la regresión lineal es un suavizador llevado al extremo). Las siguientes son algunas de las técnicas de suavizado más populares:

1. Regresión polinomial local con pesos, *lowess*.
2. Kernels.
3. Splines.

Este capítulo se centra en uso de los **splines**, ya que es la técnica de suavizado más utilizada. Los splines son funciones polinómicas a trozos de la variable explicativa, que se unen en puntos llamados **nodos**. Existen muchos tipos de splines (naturales, cílicos, B-splines, O-splines, etc.). Independientemente del tipo de spline, este capítulo se centra en los splines con penalizaciones (P-splines), que se basan en: (i) hacer una aproximación de la función $f()$ mediante una base de funciones, y (ii) añadir una penalización a la hora de estimar el modelo, de manera que se pueda controlar la variabilidad de la curva que se quiere estimar. Hay muchas maneras de representar una función a través de una base de funciones (un ejemplo sencillo de una base de funciones es el caso de la regresión polinómica, en la que la base de funciones, es decir, la matriz de regresión, es una matriz cuyas columnas son las potencias de la variable explicativa: $[X : X^2 : \dots : X^p]$). Una de las mejores opciones son los B-splines (De Boor, 2001), debido a sus buenas propiedades numéricas. La penalización se añade en la función de verosimilitud y se construye a partir de la derivada de la curva que se quiere penalizar. Generalmente se utilizan penalizaciones de orden 2, lo que implica que se penaliza todo aquello que no es lineal en la función; por tanto, si la penalización es muy grande la curva estimada es simplemente una línea recta. La penalización está controlada por un parámetro llamado **parámetro de suavizado**.

RECOMENDAR UNA O VARIAS REFERENCIAS PARA MÁS DETALLE

A la hora de ajustar este tipo de modelos hay que tomar dos decisiones importantes:

1.3. Aspectos metodológicos

13

- El número de nodos del B-spline: generalmente se utiliza esta regla:

$$\text{número de nodos} = \min\{40, \text{valores únicos de } X/4\} \quad (1.3)$$

(por ejemplo, si se tienen 100 observaciones diferentes, se elegirían $100/4=25$ nodos).

- El valor del parámetro de suavizado: se puede estimar por distintos métodos: validación cruzada, validación cruzada generalizada, **máxima verosimilitud, máxima verosimilitud restringida**, etc. Se recomienda re-expresar el modelo como un modelo mixto (véase Cap. ??) y estimarlo mediante el **método de máxima verosimilitud restringida, REML**¹, para así poder estimar el parámetro de suavizado junto con los demás parámetros del modelo.

HACE MUCHO QUE NO HAGO NADA DE ESTO Y NO TENGO TIEMPO PARA PONERME CON ELLO COMO ES DEBIDO, PERO comprobad que lo del valor del parámetro de suavizado, que las he re-escrito yo, es correcto

. La Fig. 1.1 muestra el impacto que el parámetro de suavizado tiene en el ajuste final de la curva (los datos corresponden al dataset `fossil` del paquete `Semipar`).

Por cierto, había erratas "groseras" hasta aquí; hay que releerse el texto antes de entregarlo, que si no a los demás nos quita un tiempo que no tenemos; sobre todo cuando tienes que revisar 55 capítulos dos veces

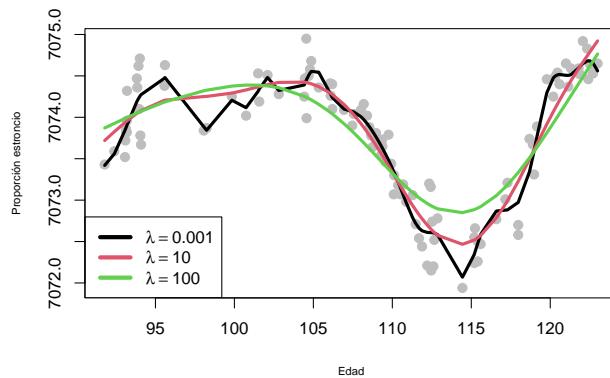


Figura 1.1: Regresión con P-splines para diferentes valores del parámetro de suavizado

1.3. Aspectos metodológicos

Al igual que en el caso de los modelos lineales y los modelos GLM, en los modelos GAM es necesario conocer algunos aspectos metodológicos que son fundamentales para llevar a cabo un

¹Es igual que el de máxima verosimilitud pero teniendo en cuenta los grados de libertad utilizados para estimar los efectos fijos al estimar los componentes de varianza (el método de máxima verosimilitud no lo hace). Para más detalles consultese [Henderson \(1953\)](#).

ajuste correcto de los modelos y entender los resultados obtenidos en el ajuste. A continuación se muestran los más relevantes.

1.3.1. Estimación de los parámetros del modelo

La estimación de los modelos GAM se lleva a cabo mediante máxima verosimilitud penalizada. Supóngase el caso de una sola variable explicativa y que se quiere ajustar el modelo:

$$Y = f(X) + \epsilon.$$

Como se comentó anteriormente, los modelos GAM tienen como punto de partida la aproximación de la función a estimar mediante una matriz formada por B-splines; es decir, se busca transformar el modelo lineal o lineal generalizado tradicional de tal forma que $f(X)$ sea el producto de una matriz multiplicada por unos coeficientes (esa matriz está formada por los B-splines). En otros términos, se elige una base (una matriz \mathbf{B}) que permita escribir la función $f(X)$ como una combinación lineal de sus elementos (los elementos de esta base son conocidos ya que se calculan a partir de las variables explicativas):

$$f(X) = \sum_{l=1}^k b_l(X)\theta_l,$$

donde $b_l(X)$ son las funciones B-spline que componen la base. En forma matricial:

$$f(X) = \mathbf{B}\theta.$$

Los parámetros θ se estiman minimizando la siguiente expresión (en el caso de datos Gaussianos los mínimos cuadrados penalizados son equivalentes a la verosimilitud penalizada):

será en caso de asumir gaussianidad para los errores y, por tanto, para la variable respuesta

$$(\mathbf{y} - \mathbf{B}\theta)'(\mathbf{y} - \mathbf{B}\theta) + \lambda\theta'\mathbf{P}\theta,$$

PONER LOS THETA EN NEGRITA NO CURSIVA

Gema dice: yo creo que no se puede, miradlo vosotros por si yo no sé

donde \mathbf{P} es la matriz de penalización y λ es el parámetro de suavizado. Dado un valor del parámetro de suavizado, las estimaciones de los parámetros vienen dadas por²:

$$\hat{\theta} = (\mathbf{B}'\mathbf{T}\mathbf{B} + \lambda\mathbf{P})^{-1}\mathbf{B}'\mathbf{y}, \quad (1.4)$$

y las estimaciones de la variable respuesta se obtienen como: $\hat{\mathbf{y}} = \underbrace{\mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{P})^{-1}\mathbf{B}'}_{\mathbf{H}}\mathbf{y}$. La matriz \mathbf{H} juega un papel importante, ya que la suma de su diagonal da una idea de la complejidad de la curva ajustada (la curva más compleja sería la que interpola los datos). Dicha suma se denomina **grados de libertad efectivos** (que no se corresponden con el número de parámetros ajustados).

²\textcolor{red}{\{Como se avanzó anteriormente, si el modelo se expresa como un modelo mixto, la estimación REML proporciona la estimación del parámetro de suavizado junto con la de los restantes parámetros del modelo}}

1.3.2. Inferencia sobre las funciones suaves

Para saber si la relación estimada entre Y y X es o no estadísticamente significativa, se debe proceder al contraste:

$$\begin{aligned} H_0 : f(X) &= 0 && \text{(no efecto)} \\ H_1 : f(X) &\neq 0 && \text{(efecto).} \end{aligned}$$

Dado que la función $f(X)$ depende de los coeficientes que acompañan a las bases de B-splines, el contraste anterior es equivalente al contraste:

$$\begin{aligned} H_0 : \boldsymbol{\theta} &= 0 \\ H_1 : \boldsymbol{\theta} &\neq 0. \end{aligned}$$

La distribución del estadístico de contraste dependerá de si la variable respuesta sigue una distribución Normal o no: en caso afirmativo el estadístico de contraste sigue un distribución F . En caso negativo, sigue una distribución χ^2 .

Comparación de modelos

Cuando se trabaja con un modelo aditivo (1.2) en el que hay más de una variable explicativa, puede ser de interés comparar versiones de ese modelo que contengan distintos conjuntos de variables. La comparación dependerá de la relación entre los modelos a comparar:

1. **Modelos anidados.** La comparación se basa, igual que en los GLM, en la diferencia en la *deviance residual*. Si se quieren comparar dos modelos m_1 y m_2 (donde $m_1 \subset m_2$), entonces:

- En el caso de variable respuesta Normal, el estadístico de contraste es:

$$\frac{(DR(m_1) - DR(m_2))/(df_2 - df_1)}{DR(m_2)/(n - df_2)} \approx F_{(df_2 - df_1), (n - df_2)},$$

donde DR es la *deviance residual* (suma de cuadrados residual) y df son los grados de libertad asociados con cada modelo.

- En otro caso, se utiliza como estadístico de contraste el siguiente:

$$DR(m_1) - DR(m_2) \approx \chi^2_{df_2 - df_1}.$$

2. **Modelos no anidados.** En este caso los contrastes anteriores no son válidos y se utilizarán criterios basados en el AIC (criterio de información de Akaike).

1.3.3. Suavizado multidimensional y para datos no Gaussianos

Para el suavizado penalizado en 2 dimensiones (o más) también se necesita una base y una penalización. El modelo sería:

$$Y = f(X_1, X_2) + \epsilon,$$

¿No ponéis término independiente, como en el caso de una sola variable?

donde $f()$ es una función de las dos covariables X_1 y X_2 . Dicha función se aproxima mediante el producto tensorial de las bases de B-splines marginales para cada una de las covariables y la penalización dependerá de dos parámetros de suavizado. Los términos de suavizado multidimensional se pueden combinar con términos unidimensionales y términos lineales. En este caso, la penalización dependería de dos parámetros de suavizado (uno para cada covariable).

La extensión de los modelos de suavizado al caso en el que la variable respuesta no sea Gaussiana, se hace de forma similar al caso lineal, cuando se pasa de un modelo de regresión lineal a un GLM. Al igual que en el caso de los GLMs, $g(\mu) = \eta = \mathbf{f}(\mathbf{X}) = \mathbf{B}\boldsymbol{\theta}$, y se añade la penalización a la función de verosimilitud de la distribución correspondiente:

$$\ell_p(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}' P \boldsymbol{\theta},$$

donde $\ell(\boldsymbol{\theta})$ es la log-verosimilitud.

1.4. Procedimiento con R: la función `gam()` del paquete `mgcv`

Aunque hay muchas librerías disponibles, la principal es `mgcv`, que implementa una gran variedad de modelos de suavizado a través de la función `gam()` (generalized additive models)³.

```
gam(formula, method="", select="", family=gaussian())
```

- `formula` es el argumento principal de esta función; es la ecuación del modelo: por ejemplo, `y ~ x1+x2+s(x3)`.
 - Lo primero que se tiene que elegir es la base a utilizar para representar las funciones suaves, `s(x)` (véase `?s` o `?smooth.terms`), o `te(x1,x2)` en el caso de suavizado bidimensional. Por defecto se usan los llamados *thin plate splines*. El tipo de base usada se puede modificar utilizando el argumento `bs` dentro de `s(x, bs = "ps")`; en este caso `ps` indica el uso de B-splines con penalizaciones. A continuación se describen otras alternativas:

³La principal referencia para esta sección es el libro de Wood (2006).

bs	Descripción
tp	Thin plate regression splines
ts	Thin plate regression splines con regularización
cr	Spline cúbicos de regresión
crs	Spline cúbicos de regresión con regularización
cc	Spline cíclicos
ps	P-splines

- **m** indica el orden de la penalización; por defecto es 2.
- **k** es el número de nodos para construir la base. El número por defecto suele ser demasiado bajo, por lo que siempre se recomienda que el usuario elija el número utilizando la regla dada en (1.3).
- **by** debe igualarse a una variable numérica o factor de la misma dimensión de cada covariable, para hacer interacciones entre curvas y variables.
- **id** se utiliza para forzar que diferentes términos suaves utilicen la misma base y la misma cantidad de suavizado.
- **method** selecciona método para estimar los parámetro de suavizado. Se puede elegir entre: **REML** (máxima verosimilitud restringida), **ML** (máxima verosimilitud), **GCV.Cp** (validación cruzada generalizada), **GACV.Cp** (validación cruzada aproximada generalizada). En la práctica, como se indicó anteriormente, se prefiere **REML**.
- **family** permite elegir la distribución de la variable respuesta (binomial, Poisson, etc.); por defecto asume Gaussiana.
- **select=TRUE** contrasta si una variable debe entrar o no en el modelo.

PREGUNTA TONTA: ¿Si en **method** se elige **REML**, entonces se transforma previamente el modelo en un modelo mixto para sacar las estimaciones?

1.5. Casos prácticos

En este apartado se ven una serie de aplicaciones que permiten mostrar los diferentes usos de este tipo de modelos.

1.5.1. Modelo unidimensional con **fossil**

Se empieza ilustrando el uso de la función **gam()** con el conjunto de datos **fossil** del paquete **SemiPar**. El objetivo es estimar la relación entre la edad de los fósiles y la proporción de isotopos de estroncio.

```
library("SemiPar")
data(fossil)
Y <- 1000*fossil$strontium.ratio
X <- fossil$age
plot(X,Y, xlab="Edad", ylab = "Proporción de estroncio")
```

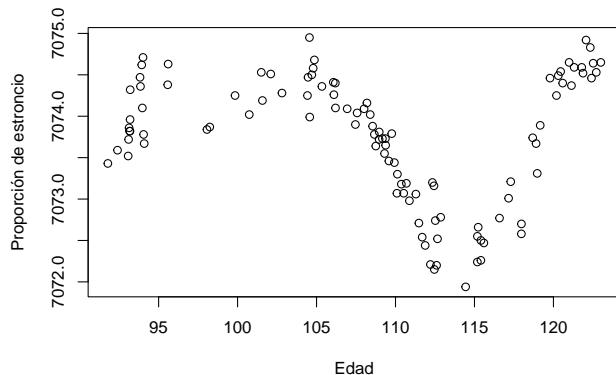


Figura 1.2: Edad de los fósiles con respecto a la proporción de isótopos de estroncio

A la vista de la Fig. 1.2, está claro que se necesita ajustar una curva (y no una línea) para estimar la relación entre ambas variables. Para ello se utiliza la función `gam()`, que devuelve un objeto de tipo "gam" y que se puede usar con las típicas funciones `print()`, `summary()`, `fitted()`, `plot()`, `residuals()`, etc.

```
library("mgcv")
fit_gam <- gam(Y ~ s(X, k=25, bs="ps"), method="REML", select=TRUE)
# se eligen 25 nodos ya que se lavariable tiene 106 observaciones
summary(fit_gam)
#>
#> Family: gaussian
#> Link function: identity
#>
#> Formula:
#> Y ~ s(X, k = 25, bs = "ps")
#>
#> Parametric coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 7.074e+03 2.435e-02 290504    <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Approximate significance of smooth terms:
#>             edf Ref.df   F p-value
#> s(X) 10.22     24 35.89    <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> R-sq.(adj) =  0.891  Deviance explained = 90.2%
#> -REML = 23.946  Scale est. = 0.062849  n = 106
```

1.5. Casos prácticos

19

Como se puede ver, la relación entre la variable respuesta (Y , proporción de estroncio) y la variable explicativa (X , edad) se ha especificado mediante un *spline*, `s()`, de tipo penalizado, `ps`, con 25 nodos. **Se ha seleccionado ‘REML’ como método para estimar el parámetro de suavizado (los parámetros del spline se estiman también mediante ‘REML’, ya que da lugar a las mismas estimaciones que máxima verosimilitud).**

En la primera parte de la salida anterior aparecen los términos que entran linealmente en el modelo (en este caso sólo el término independiente o intercepto); en la parte de abajo se muestran los términos de suavizado. Como se indicó anteriormente, dado que se ha usado `select=TRUE`, se está contrastando si la variable `edad` debe entrar en el modelo o no. En este caso, es claro que ha de entrar ya que el p-valor de `s(x)` es pequeño y los grados de libertad asociados son aproximadamente 10, lo que indica que la relación entre Y y X está lejos de la linealidad.

habría que comentar el papel que juegan los grados de libertad asociados para decidir si una variable entra o no en el modelo

La función `gam.check()` devuelve los gráficos de residuos usuales (residuos frente a valores ajustados, gráficos de cuantiles para comprobar la normalidad, etc.), pero además proporciona información sobre el proceso de ajuste del modelo.

```
gam.check(fit_gam, cex=1.2)
```

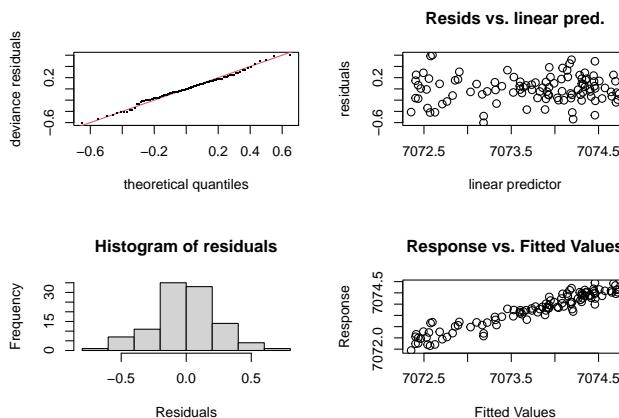


Figura 1.3: Gráficos de residuos obtenidos con ‘`gam.check()`’

```
#>
#> Method: REML    Optimizer: outer newton
#> full convergence after 5 iterations.
#> Gradient range [-4.557319e-06,5.900236e-06]
#> (score 23.94602 & scale 0.06284944).
#> Hessian positive definite, eigenvalue range [4.557347e-06,53.03185].
```

```
#> Model rank = 25 / 25
#>
#> 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(X) 24.0 10.2     1.03    0.58
```

El test que aparece en la parte de abajo está contrastando si el número de nodos elegido es suficiente. Si el valor de k está muy próximo al de edf , entonces se debería reajustar el modelo con más nodos.

El comando `plot()` permite dibujar la función suave que relaciona Y con X. La curva estimada que aparece en la Fig. 1.4 está centrada (la función `plot()` siempre lo hace de esta forma), el argumento `shade` hace que se sombre el intervalo de confianza y `seWithMean` hace que la incertidumbre sobre el término independiente se incluya en el cálculo del intervalo de confianza.

```
plot(fit_gam, shade=TRUE, seWithMean=TRUE, pch=19, 1, cex=.55)
```

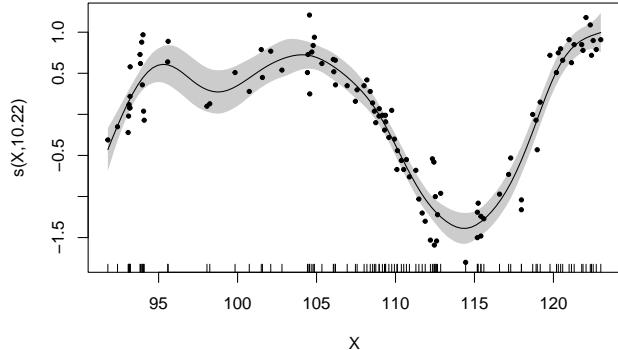


Figura 1.4: Curva ajustada e intervalo de confianza

1.5.2. Modelo aditivo con airquality

En esta sección se analizan de nuevo los datos `airquality` (ver `airquality`⁴), que consisten en 154 medidas de calidad del aire en Nueva York, de mayo a septiembre 1973. El objetivo es establecer la relación entre las variables meteorológicas y el nivel de concentración de ozono en la atmósfera. Ya se ha analizado dicha relación en el Cap. ??, donde los ajustes lineales realizados eran

⁴Conjunto de datos incluido con la instalación base de R.

1.5. Casos prácticos

21

satisfactorios pero se encontraban problemas en los residuos del modelo, lo cual impedía validar la modelización realizada. Allí se sugería que la relación entre la variable respuesta y alguna explicativa fuese no lineal. Además, se consideró la transformación logarítmica de la variable `Ozone`, y con dicha trasformación se obtenía una distribución más similar a la distribución Normal.

En consecuencia, se va a ajustar el modelo incluyendo las variables explicativas sin imponerles linealidad; en particular, se van a incluir las variables `Wind`, `Temp` y `Solar.R`. Las variables `Wind` y `Temp` tienen sólo 31 y 40 valores únicos, respectivamente, aunque el conjunto de datos tiene 154 valores; por eso, para estas dos variables, se decide establecer el número de nodos en 10 y no más; para la variable `Solar.R` el número de nodos se fija en 20.

```
airq_gam=gam(log(Ozone)~s(Wind,bs="ps",k=10) +
               s(Temp,bs="ps",k=10)+s(Solar.R,bs="ps",k=20),
               method="REML",select=TRUE,data=airquality,na.action=na.omit)
summary(airq_gam)
#>
#> Family: gaussian
#> Link function: identity
#>
#> Formula:
#> log(Ozone) ~ s(Wind, bs = "ps", k = 10) + s(Temp, bs = "ps",
#>       k = 10) + s(Solar.R, bs = "ps", k = 20)
#>
#> Parametric coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 3.41593   0.04586  74.49 <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Approximate significance of smooth terms:
#>             edf Ref.df    F p-value
#> s(Wind)     2.318     9 2.255 3.13e-05 ***
#> s(Temp)      1.852     9 6.128 < 2e-16 ***
#> s(Solar.R)  2.145    19 1.397 2.31e-06 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> R-sq.(adj) =  0.689  Deviance explained = 70.7%
#> -REML = 86.106  Scale est. = 0.23342 n = 111
```

Los resultados indican que todas las variables son significativas (p-valores pequeños), estando la variable `Temp` próxima a la linealidad (los grados de libertad efectivos asociados a la variable son 1.8). El R^2 ajustado es 0.69, por lo que el modelo ajusta moderadamente bien los datos. La Fig. 1.5 muestra las tres curvas ajustadas, incluyendo los llamados *residuos parciales* que corresponden a, por ejemplo, en el caso del gráfico del viento, $\log(Ozone) - \hat{\beta}_0 - \hat{f}(Temp) - \hat{f}(Solar.R)$, es decir, lo que queda sin explicar después de haber ajustado los demás términos del modelo.

```
library("mgcv")
b <- getViz(airq_gam)
# getViz es otra opción para dibujar los términos de un modelo gam()
print(plot(b, allTerms = T, shade=T), pages = 1)
```

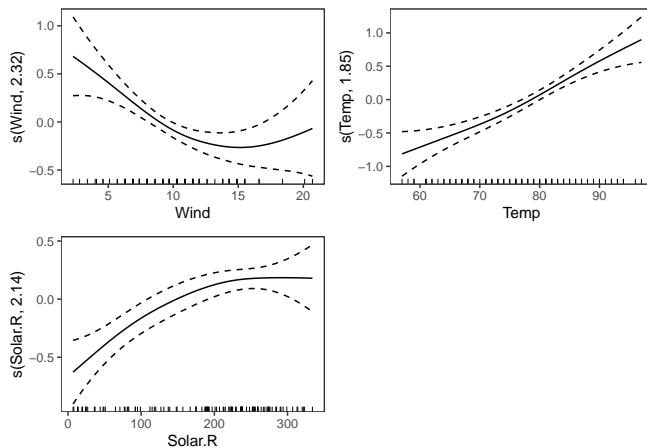


Figura 1.5: Curvas estimadas para ‘Wind’, ‘Temp’ y ‘Solar’

para cerrar el círculo, tenéis que dar una predicción y enseñar cómo se conforma esa predicción a partir de las curvas de la Fig. 1.5 y el término independiente

perdonad, quizás sea cansancio: ¿dónde están los residuos parciales en las figuras. Lo que hay es el spline y un intervalo de confianza no?

1.5.3. Modelo semiparamétrico con onions

Es un caso particular del modelo aditivo, pues en este modelo todas las variables entran de forma lineal excepto una:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1} + f(X_p) + \epsilon.$$

La forma de ajustar el modelo es exactamente igual a la anterior. Pero hay un caso que merece especial interés: cuando en la parte paramétrica se incluye una variable categórica con dos o más niveles. Al igual que en el caso de regresión lineal, se puede plantear si se quieren ajustar dos o más rectas paralelas (modelo aditivo) o no paralelas (modelo con interacción).

Para ilustrar este caso se acude al `data.frame onions` (librería `SemiPar`). Contiene 84 observaciones de un experimento sobre la producción de un tipo de cebolla en dos localidades: (Purnong Landing (la localidad de referencia) y Virginia. El objetivo es relacionar el logaritmo de la producción de cebollas con la densidad de plantas por metro cuadrado, `dens`. El modelo lineal básico sería:

$$\log(\text{yield}_j) = \beta_0 + \beta_1 \text{location}_j + \beta_2 \text{dens}_j + \epsilon_j$$

donde

$$\text{location}_j = \begin{cases} 0 & \text{si la observación } j \text{ es de Purnong Landing} \\ 1 & \text{si la observación } j \text{ es de Virginia} \end{cases}$$

Se comienza por ajustar el siguiente modelo:

$$\log(\text{yield}_j) = \beta_0 + \beta_1 \text{location}_j + f(\text{dens}_j) + \epsilon_j$$

```
library("mgcv")
library("SemiPar")
data(onions)

#Se indica a R que la variable locationVirginia es categórica
onions$location <- factor(onions$location)
#Se recodifica la variable
levels(onions$location) <- c("Purnong Landing", "Virginia")
fit1 <- gam(log(yield) ~ location + s(dens, k=20, bs="ps"),
            method="REML", select=TRUE, data=onions)
summary(fit1)
#>
#> Family: gaussian
#> Link function: identity
#>
#> Formula:
#> log(yield) ~ location + s(dens, k = 20, bs = "ps")
#>
#> Parametric coefficients:
#>                               Estimate Std. Error t value Pr(>|t|)
#> (Intercept)           4.85011   0.01688 287.39    <2e-16 ***
#> locationVirginia -0.33284   0.02409  -13.82    <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Approximate significance of smooth terms:
#>          edf Ref.df   F p-value
#> s(dens) 4.568     19 72.76    <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> R-sq.(adj) =  0.946  Deviance explained = 94.9%
#> -REML = -54.242  Scale est. = 0.011737 n = 84
```

En este ejemplo se ve que en la parte lineal aparecen dos parámetros, ambos significativos: la ordenada en el origen o intercepto y el coeficiente de la categoría **Virginia** de la variable **location**, que es negativo, indicando que la producción media en Purnong Landing es mayor que en Virginia. El término de suavizado también es significativo.

En este caso, función `plot.gam()` sólo dibuja una curva, pues las curvas para las dos localizaciones son paralelas y la diferencia entre ellas es igual al valor del parámetro correspondiente a **localización**. Para dibujar las curvas para cada localización se utiliza la función

`plot_smooth()` de la librería `tidymv`. Los argumentos son, primero el modelo, después la variable explicativa y por último la variable categórica.

```
library("tidymv")
library("ggplot2")
plot_smooths(fit1, dens, location) +
  theme(text = element_text(size = 12))
```

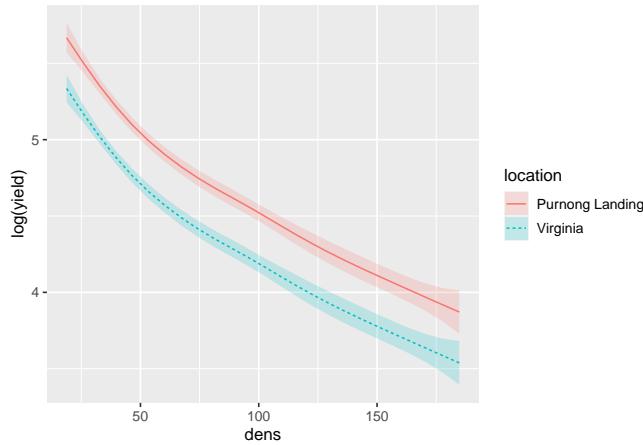


Figura 1.6: curvas ajustadas para ambas localidades

Asumir curvas paralelas para ambas localidades implica que el descenso en la producción de cebollas a medida que aumenta la densidad de plantas es el mismo para ambas localidades, y esto no tiene por qué ser cierto. Para relajar esta hipótesis se puede ajustar un modelo con interacción (de manera similar a lo que se hace en el caso de regresión lineal):

$$\log(\text{yield}_j) = \beta_0 + \beta_1 \text{location}_j + f(\text{dens}_j)L(j) + \epsilon_j$$

donde

$$L(j) = \begin{cases} 0 & \text{si la } j\text{-ésima observación es de Purnong Landing} \\ 1 & \text{si la } j\text{-ésima observación es de Virginia} \end{cases}$$

Para hacerlo en **R**, se introduce el argumento `by=location` dentro de la curva

```
fit2 <- gam(log(yield) ~ location + s(dens, k=20, bs="ps", by=location),
             method="REML", data=onions)
summary(fit2)
#>
#> Family: gaussian
#> Link function: identity
#>
#> Formula:
#> log(yield) ~ location + s(dens, k = 20, bs = "ps", by = location)
```

1.5. Casos prácticos

25

```
#>
#> Parametric coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 4.84415   0.01603 302.19 <2e-16 ***
#> locationVirginia -0.33018   0.02270 -14.54 <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Approximate significance of smooth terms:
#>          edf Ref.df     F p-value
#> s(dens):locationPurnong Landing 3.096 3.834 176.9 <2e-16 ***
#> s(dens):locationVirginia       4.742 5.795 153.0 <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> R-sq.(adj) = 0.952 Deviance explained = 95.7%
#> -REML = -58.541 Scale est. = 0.010446 n = 84
```

Ahora aparecen dos términos suaves, uno para cada localidad, de modo que estas curvas no tienen por qué ser paralelas, sino que cada una se ajustará a la forma que tengan los datos. En este caso, la Fig. 1.7, generada de nuevo con `plot_smooths`, muestra como las curvas se van alejando a medida que aumenta la densidad de plantas.

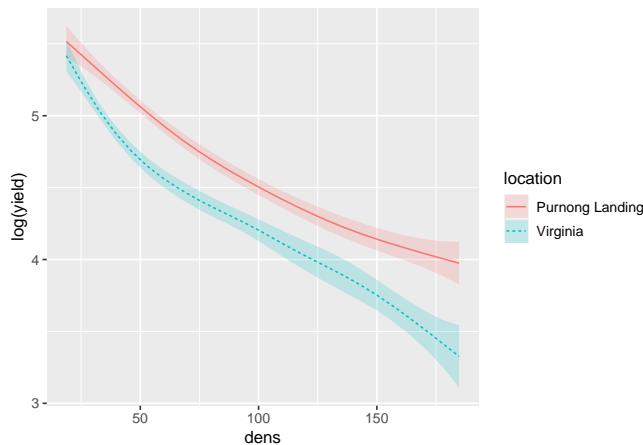


Figura 1.7: Curvas ajustadas para ambas localidades permitiendo que no sean paralelas

Para finalizar se comparan ambos modelos con el criterio AIC.

```
AIC(fit1); AIC(fit2)
#> [1] -125.2307
#> [1] -131.2181
```

Dado que el menor valor se alcanza en el segundo modelo, se escogería el modelo que incluye la interacción entre la variable densidad y la localidad.

1.5.4. Modelo aditivo generalizado y multidimensional con `smacker`

En este epígrafe se analizan los datos `smacker` del paquete `sm`. El objetivo es ver cómo influyen las condiciones del mar (temperatura de agua, etc.) en la ausencia o presencia de huevos de jurel en el mar Cantábrico. Además, se incorporará al modelo la posición geográfica mediante las covariables latitud y longitud; de esta forma se podrá captar el efecto espacial.

```
library("sm")
data(smacker)
library("dplyr")
smacker <- smacker |>
  mutate(Presence = ifelse(Density>0, 1, 0),
         smack.long = -smack.long,
         ldepth = log(smack.depth))
library("maps")
par(pty="s")
Position <- cbind(smacker$smack.long, smacker$smack.lat)
plot(Position, col=NULL, xlim=c(-10,-1), ylim=c(43,48), cex=1.2, xlab="longitud",
      ylab="latitud")
map("world", add=TRUE, fill=TRUE, col="grey")
points(Position[smacker$Presence==1,], pch=1, cex=.5, col=4)
points(Position[smacker$Presence==0,], pch=16, cex=.5, col=2)
legend("topleft", c("Presencia ", "Ausencia"), col=c(4,2), pch=c(1,16), cex=.85)
```

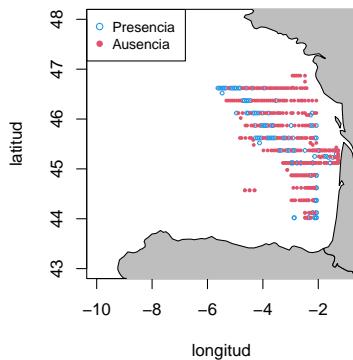


Figura 1.8: Área donde se constató la ausencia/presencia de huevos de jurel

Dado que la variable respuesta es dicotómica, se utiliza un modelo de regresión logística en el que se flexibiliza la relación lineal de las variables explicativas con la respuesta y, además, se

1.5. Casos prácticos

27

usa una superficie para estimar el efecto de la localización como una función en dos dimensiones (latitud y longitud). En este caso, en vez de usar `te()` se usa `s()` también para el caso de 2 dimensiones. La diferencia fundamental con `te()` es que `s()` asume un suavizado isotrópico, es decir, el mismo parámetro de suavizado para la latitud y longitud. No se debe usar `s()` para el suavizado en dos dimensiones si las covariables están medidas en unidades diferentes (en este caso como `no` lo están, se puede usar el suavizado isotrópico):

```
logit1 <- gam(Presence~s(ldepth)+ s(Temperature)+ s(smack.long, smack.lat,k=60),
                family=binomial, select=TRUE, data=smacker)
b <- getViz(logit1)
print(plot(b, allTerms = T), pages = 1)
```

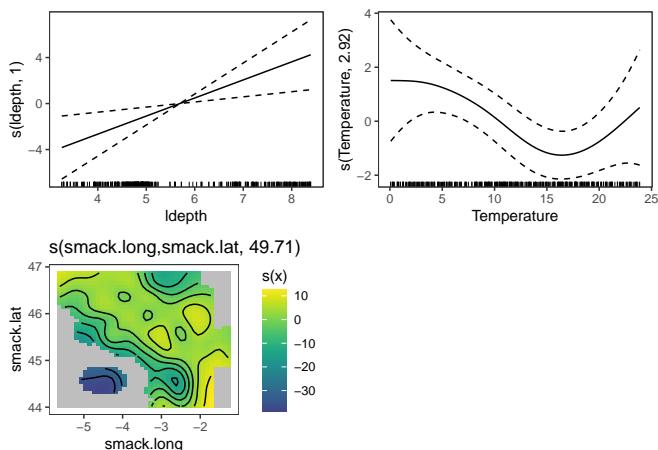


Figura 1.9: Efectos suaves estimados por el modelo para las variables. Efecto de la profundidad y temperatura en la fila superior y efecto espacial en la inferior

En la Fig. 1.9 se aprecia que la relación entre la probabilidad de presencia de huevos y la temperatura no es lineal, mientras que sí lo es en el caso de la profundidad. El R^2 es tan sólo 0,4, por lo que convendría utilizar más variables explicativas para obtener buenas predicciones.

Las probabilidades predichas se pueden obtener con la función `predict`.

```
prob=predict(logit1,type="response")
```

Resumen

En este capítulo se introducen los modelos aditivos generalizados. En particular:

- Se presentan distintos aspectos metodológicos de carácter inferencial en este tipo de modelos.
- Se muestra el uso de R para llevar a cabo su ajuste.
- Se presentan diversos casos prácticos que ilustran la versatilidad de estos modelos para analizar datos complejos.

Bibliografía

De Boor, C. (2001). *A practical guide to splines*. Applied Mathematical Sciences. Springer-Verlag, New York.

Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics*, 9:226–252.

Wood, S. N. (2006). *Generalized Additive Models - An introduction with R*. Texts in Statistical Science. Chapman & Hall.

Índice alfabético

base, 14
deviance, 15
grados de libertad efectivos, 14
modelo
 aditivo, 12
parámetro de suavizado, 12
residuos
 parciales, 21
spline, 19