La importancia de los modelos multidimensionales en el campo de la epidemiología

Pedro Saavedra Santana
Catedrático de Estadística, Investigación Operativa y Computación
Departamento de Matemáticas, Universidad de Las Palmas de Gran Canaria

1. Introducción

Evaluar el efecto de una determinada exposición sobre una variable de respuesta (*outcome*) es uno de los objetivos más frecuentes de los estudios epidemiológicos. Así, por ejemplo, para explorar la asociación de una cierta *exposición* con una patología, pueden compararse las prevalencias de ésta en las cohortes definidas por la existencia o no de tal exposición. Si se desea analizar su influencia sobre una variable numérica, cabe comparar entonces sus valores medios correspondientes a las referidas cohortes. Este procedimiento, en realidad, sólo debe constituir una primera aproximación al problema, pues generalmente las diversas cohortes no son a menudo comparables por el efecto de la *confusión* de otras variables. Es bien sabido que la masa ósea de cualquier sujeto alcanza su máximo valor alrededor de los 30 años y a partir de esa edad se produce un suave declive que en la mujer se acelera a partir de la menopausia. A su vez, la hipertensión arterial (HTA) suele progresar a partir de ciertas edades. Por tal motivo, es posible que en la cohorte de personas con HTA, la masa ósea sea significativamente menor que en el grupo de normotensos. Esto no significa que la HTA sea causa de la osteoporosis (fragilidad ósea), pues la depleción de la masa ósea en los hipertensos se explicaría por el hecho de que éstos habitualmente son personas de mayor edad que los normotensos. Para investigar si la HTA es causa de osteoporosis habría al menos que determinar si a una *misma edad*, la osteoporosis tiene mayor prevalencia en los hipertensos. La edad actúa como factor de confusión, pero no podemos descartar que puedan existir otros factores de esta naturaleza. Es, por tanto, conveniente hacer las comparaciones *ajustando* por todos estos factores. Generalmente, estos ajustes se hacen en el contexto de los modelos de análisis de la covarianza (ajustes lineales). Sin embargo, cuando el factor de confusión influye de forma no lineal sobre el *outcome*, las comparaciones ajustadas podrían ser erróneas. En este sentido, los ajustes no paramétricos podrían ser ventajosos. Una alternativa interesante en el campo de la epidemiología son los modelos aditivos generalizados (GAM), en los que los efectos no lineales se ajustan mediante splines cúbicos o estimadores de núcleo.

En el tercer epígrafe de este trabajo se analiza la asociación entre una variable de respuesta (*outcome*) y otra explicativa, la cual a menudo representa una cierta exposición. Se presentan los métodos de regresión lineal y alternativas no paramétricas tales como los estimadores de núcleo y los splines cúbicos. Con el objetivo esencial de ajustar por posibles variables de confusión se introducen los modelos de regresión lineal múltiple y los modelos aditivos generalizados. Estos últimos se analizan en el contexto no paramétrico. Los métodos

expuestos se ilustran con dos conjuntos de datos simulados y otro correspondiente a un estudio de diabetes realizado en el municipio de Telde (Gran Canaria).

2. Datos

Ilustramos los métodos desarrollados mediante tres conjuntos de datos. El primero corresponde a un estudio sobre diabetes mellitus tipo 2 (DM2) realizado en el municipio de Telde (Gran Canaria). El objetivo de este estudio consistía en estimar las prevalencias de diabetes en la población mayor de 30 años (cruda y ajustada por poblaciones estándar), así como analizar los factores asociados con esta patología. El estudio es de diseño transversal y para su realización se seleccionó una muestra aleatoria de la referida población, lográndose finalmente un total de 1030 personas evaluables.

En orden a valorar de forma precisa la bondad de ajuste de los modelos considerados, se consideró también un conjunto de datos simulados a partir de tres predictores. El conjunto de datos (SIMU-1) se obtuvo como:

$$Y_i = 4 + 2x_{i,1} + 5\cdot\left(x_{2,i} - 5\right)^2 + 10\cdot\cos\left(3\cdot x_{3,i}\right) + e_i \ : \ i = 1,\cdots,n \ ,$$

siendo e_i variables independientes y distribuidas $N\left(0,\sigma = .3\right)$. Este modelo es aditivo pero con dos componentes no lineales. Esto permitirá ver la ventaja de los métodos no lineales frente al habitual modelo de ajuste lineal.

Se simularon también datos basados en la evolución del patrón de masa ósea que se describen en el cuarto epígrafe.

3. Modelos de regresión con un predictor

En esta sección haremos una breve revisión de los modelos lineales, tanto para respuesta numérica como binaria. Inicialmente tratamos los modelos con un único predictor. En tal caso, la exploración de modelación no lineal puede hacerse fácilmente utilizando las técnicas de suavizamiento no paramétricas, tales como los estimadores de núcleo o los splines cúbicos. En el caso multidimensional, la exploración no lineal es más compleja. En este trabajo consideramos los modelos aditivos generalizados (GAM) estimados a través de los algoritmos *backfitting* (tanto para respuesta numérica como categórica).

3.1. Regresión lineal con un solo predictor. Considérese que condicionalmente a cada valor x_i de una variable X se observa una variable aleatoria Y_i de tal forma que $Y_1,\cdots,Y_n$ son independientes. Esta hipótesis resulta natural cuando cada observación $\left(x_i, Y_i\right)$ se realiza sobre unidades muestrales independientes. Para el caso en el que las variables Y_i sean numéricas, el modelo de regresión lineal supone que el valor esperado tiene la forma:

(RLinS) $$\mu\left(x_i\right) = E\left[Y_i \middle| X = x_i\right] = \alpha + \beta\cdot x_i \ .$$

Este ajuste lineal permite disponer de una simple interpretación del parámetro β, el cual representa la variación en el valor esperado de la variable Y_i por unidad de variación del

predictor X. El modelo clásico de regresión lineal simple supone que las variables aleatorias Y_i son independientes con distribución de probabilidad $N\left(\mu(x_i),\sigma\right)$, para $i=1,\cdots,n$. Esta hipótesis permite disponer de una función de verosimilitud a través de la cual pueden estimarse los parámetros α, β y σ.

En los estudios biomédicos es muy frecuente que la variable de respuesta Y_i sea binaria, indicando la presencia o ausencia de un carácter. Codificamos la presencia por (1) y la ausencia por (0). En este caso, el modelo (RLinS) obviamente no es adecuado. Alternativamente, puede considerarse el modelo de regresión logística de la forma:

(RLogS) $$\operatorname{logit}\left\{\mathbb{P}\left(Y_i=1\middle|X=x_i\right)\right\}=\alpha+\beta\cdot x_i\,,$$

siendo la función $\operatorname{logit}(z)=\log\left(z/(1-z)\right)$. El modelo (RLogS) puede alternativamente expresarse por:

$$\pi(x_i)=\mathbb{P}\left(Y_i=1\middle|X=x_i\right)=\frac{\exp(\alpha+\beta\cdot x_i)}{1+\exp(\alpha+\beta\cdot x_i)}\,.$$

De esta forma, puede afirmarse que las variables aleatorias $Y_1,\cdots,Y_n$ son independientes y con distribución de probabilidad común $b\left(1,\pi(x_i)\right)$ (***binomial***). Al igual que en los modelos (RLinS), el parámetro β tiene una interpretación de gran utilidad en los estudios biomédicos, indicando la variación del riesgo de la presencia del evento o carácter ($Y=1$) por unidad de variación del predictor. Es muy directo deducir la siguiente expresión:

$$\exp(\beta)=\frac{\mathbb{P}\left(Y=1\middle|X=x+1\right)\Big/\mathbb{P}\left(Y=0\middle|X=x+1\right)}{\mathbb{P}\left(Y=1\middle|X=x\right)\Big/\mathbb{P}\left(Y=0\middle|X=x\right)}\,.$$

Esta expresión recibe el nombre de *odd-ratio* y evalúa el aumento de riesgo del evento $\{Y=1\}$ por unidad de incremento del predictor X. Supóngase por ejemplo que X es a su vez una variable binaria indicatriz de un exposición (1=*sí*; 0=*no*). Si el evento modelado por Y es una determinada enfermedad, el numerador de la *odd-ratio* expresa el número de enfermos por cada no enfermo en la cohorte de sujetos expuestos al factor predictor ($X=1$), mientras que el denominador es la misma relación en la cohorte de sujetos no expuestos ($X=0$).

3.2. Regresión no paramétrica con un predictor. La representación lineal de $\mu(x)=E\left[Y\middle|X=x\right]$ es habitualmente conveniente debido a su fácil interpretación, y además, porque es la aproximación de Taylor de primer orden a $\mu(x)$. No obstante, el epidemiólogo requiere a menudo conocer de forma más precisa el modo en el que la variable X se asocia con la respuesta. En algunos casos se conoce una forma paramétrica precisa de $\mu(x)$, y a

través de los datos observados puede darse el ajuste adecuado para el modelo utilizando los métodos de regresión no lineal. Sin embargo, cuando nos encontramos ante un nuevo problema, raramente se dispone de una forma paramétrica para $\mu(x)$. En tal caso, los métodos de alisamiento proporcionan una herramienta eficaz para explorar la forma de $\mu(x)$. Los métodos de alisamiento de datos más usuales son los estimadores de núcleo y los splines cúbicos.

Supóngase entonces que el conjunto de datos de la forma $\{(x_i, Y_i) : i = 1, \cdots, n\}$ obedece a un modelo de la forma $Y_i = \mu(x_i) + e_i$, siendo $E[e_i] = 0$ y $\text{var}(e_i) = \sigma^2$. Un estimador básico de núcleo tiene la forma:

(K) $$\hat{\mu}_\lambda(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

donde $K(z)$ es la función núcleo que satisface $\int_{-\infty}^{\infty} K(z) \cdot dz = 1$ y h es el ancho de ventana (*bandwidth*). Nótese además que el orden de diferenciabilidad del estimador $\hat{\mu}_h(x)$ es el mismo que el de la función de núcleo $K(z)$.

La elección particular de la función de núcleo no tiene especial influencia en el estimador $\hat{\mu}_h(x)$. El ancho de ventana h se corresponde con el grado de alisamiento y su selección es crucial en la estimación (ver Härdle, 1991, Cap. 4).

El ajuste por splines cúbicos es un caso especial de modelación basada en un conjunto de funciones *base* $b_j(x) : j = 1, \cdots, p$, plenamente identificadas. Más concretamente, la función de medias es de la forma $\mu(x) = \sum_{j=1}^{p} \alpha_j \cdot b_j(x)$. Elecciones clásicas son las funciones polinómicas ($b_j(x) = x^{j-1}$). Las funciones base de los splines cúbicos se definen a partir de un conjunto de nodos $\{\xi_1, \cdots, \xi_K\}$ localizados en el interior del rango de la variable x, de tal forma que en cada intervalo $]\xi_{j-1}, \xi_j[$, $\mu(x)$ coincide con un polinomio de tercer grado, y en los nodos sus dos primeras derivadas son continuas. Una base elemental de splines cúbicos está formada por las funciones $b_j(x) = |x - \xi_j|^3$ para $j = 1, \cdots, K$, $b_{K+1}(x) = 1$ y $b_{K+2}(x) = x$.

El problema esencial de ajuste por splines cúbicos es la elección del conjunto de nodos, la cual influye notablemente en los resultados del ajuste. Nótese que $K = 0$ supone un ajuste lineal, mientras que valores de K próximos a n conducen a una situación de sobre-ajuste. Este problema puede obviarse mediante un planteamiento alternativo del ajuste. Tal planteamiento se basa en los residuales cuadráticos penalizados. Más concretamente, se trata de estimar $\mu(x)$ como la función que minimiza la suma de residuales cuadráticos penalizados:

$$PRSS(\mu,\lambda)=\sum_{i=1}^{n}\left\{\ Y_i-\mu(x_i)\ \right\}^2+\lambda\int\left\{\mu''(x)\right\}^2dx\ .$$

El primer término mide la proximidad de los datos a $\mu(x)$, mientras que el segundo penaliza la curvatura de la función. La elección $\lambda=0$ supone que el estimador es una interpolación de los datos, mientras que $\lambda=\infty$ conduce a tomar $\mu''(x)=0$, y de esta forma el ajuste es lineal. Puede probarse que el problema de cálculo variacional $\min_{\mu} R(\mu,\lambda)$ tiene una única solución, la cual consiste en un spline natural cúbico (ver apéndice) cuyo conjunto de nodos es el de valores de la variable X, $\{x_i : i=1,\cdots,n\}$. El problema del número de nodos se traslada ahora al problema de selección del parámetro de suavizamiento λ. Este parámetro puede también estimarse mediante métodos de validación cruzada (ver Hastie y Tibshirani, 1990, p. 42).

Para respuesta binaria, el ajuste no paramétrico alternativo al modelo (RLogS) tiene la forma:

$$\text{logit}\left\{\mathbb{P}\left(Y_i=1\middle|X=x_i\right)\right\}=f(x_i)$$

donde también supondremos que la función $f(x)$ es suave en algún sentido. El modelo anterior puede alternativamente expresarse por:

$$\pi(x_i)=\mathbb{P}\left(Y_i=1\middle|X=x_i\right)=\frac{\exp\left(f(x_i)\right)}{1+\exp\left(f(x_i)\right)}.$$

Si suponemos que la función $f(x)$ es dos veces diferenciable, puede entonces estimarse a través de la siguiente función de verosimilitud penalizada:

$$\ell(f;\lambda)=\sum_{i=1}^{n}\left[Y_i\log\pi(x_i)+(1-Y_i)\log\left(1-\pi(x_i)\right)\right]-\lambda\int\left\{f''(x)\right\}^2dx\ .$$

El primer término se corresponde con la log-verosimilitud basada en la distribución binomial. El estimador aquí considerado es la función $\hat{f}(x)$ solución del problema variacional $\max_f \ell(f;\lambda)$. Puede aquí también probarse que este problema tiene una única solución, la cual es un spline natural cúbico con nodos en el conjunto de valores $\{x_i : i=1,\cdots,n\}$.

Ilustramos ahora los métodos de alisamiento para variable de respuesta numérica utilizando el conjunto de datos $x_i=i/100, i=1,\cdots,100$, $Y_i=\cos(5\cdot x_i)+e_i$, siendo e_i variables aleatorias independientes y con distribución $N(0,\sigma=.3)$. La Figura 1.a muestra la

función de esperanzas $\mu(x)$, los datos y ajustes de núcleo y por splines cúbicos. Para variable de respuesta binaria consideramos los datos de diabetes descritos en el segundo epígrafe. Más concretamente, ajustamos la probabilidad de diabetes mellitus en función de la edad mediante un modelo de regresión logística no lineal en el cual la función $f(x)$ se estima mediante la función de verosimilitud penalizada.

(a) (b)

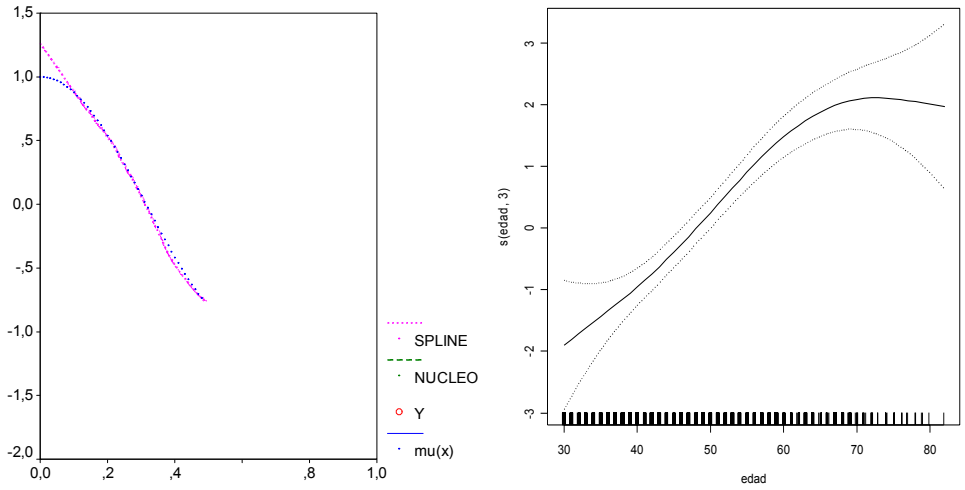

Figura 1. (a) Representación simultánea de los estimadores de núcleo y spline cúbico.
(b) Evolución del riesgo de diabetes mellitus en función de la edad. Ajuste por un spline natural cúbico.

4. Modelos de regresión múltiple

Los modelos de regresión con un solo predictor en epidemiología son por sí mismos de escaso interés. La asociación entre predictor y respuesta que puede establecerse a través de tales modelos puede raramente traducirse en una relación de causalidad. La Figura 2 muestra una simulación de la evolución de la masa ósea media determinada por DXA en una determinada población. Está establecido que ésta alcanza su valor máximo alrededor de la edad de 30 años, iniciándose luego un declive que en la mujer se acelera en el periodo de la menopausia. En el gráfico se distingue la determinación DXA en dos grupos de mujeres, a saber: mujeres asmáticas tratadas con esteroides (A) y mujeres control "sanas" (C). El estudio de simulación se basa en un diseño *descuidado* en el que las mujeres asmáticas son claramente de mayor edad que los controles. Es habitual en los estudios biomédicos que dos o más cohortes difieran en edad, aunque no con la exageración de este estudio.

Obviamente la media de la masa ósea en el grupo de mujeres asmáticas tratadas con esteroides es inferior a la del grupo control, pero esta diferencia no es atribuible al tratamiento o a la enfermedad sino a la edad. Un modelo que explique la masa ósea sólo a través de la variable de grupo (modelo ANOVA) pondría de manifiesto que la masa ósea media en los controles ($\mu_C = 0.927$) es superior a la media en el grupo de asmáticas ($\mu_A = 0.795$) de

forma estadísticamente significativa ($p < 0.001$). Ello desde luego tiene escaso interés epidemiológico, dado que esta diferencia es atribuible exclusivamente al efecto de *confusión* de la edad. El interés desde el punto de vista epidemiológico radica precisamente en evaluar el efecto real de la *exposición* analizada.

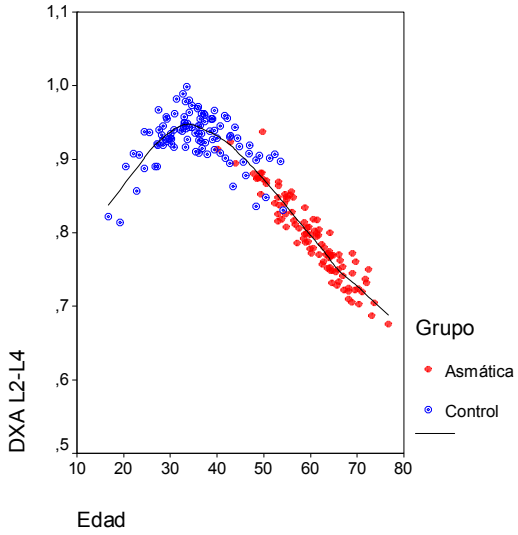

Figura 2. Simulación de la evolución de la masa ósea (DXA en L2-L4) a lo largo de la edad y según cohorte definida por la presencia o no de enfermedad asmática tratada con esteroides.

Una forma más interesante de explicar la DXA es a través de un modelo que incluya las variables *edad* y grupo (g). Sea:

$$\mu(g,edad) = \theta + \alpha_g + \beta \cdot (edad) \; : \; g = A,C \; ; \; \alpha_A = \alpha \; ; \; \alpha_C = 0 \, .$$

Aquí, $\mu(g,edad)$ representa el valor esperado de la masa ósea (DXA) en el grupo g (A ó C) para cada valor de *edad* especificado. Este modelo supone que existe una diferencia constante α a lo largo de todas las edades entre los valores esperados de la DXA en asmáticas y controles (no interacción). El modelo supone también que en cada uno de los grupos de estudio el efecto de la edad sobre el valor esperado de la DXA es lineal. Esta última hipótesis está en clara contradicción con la Figura 2, en la que se muestra que el efecto de la edad es claramente no lineal. Una alternativa más verosímil al modelo anterior sería:

$$\mu(g,edad) = \theta + \alpha_g + f(edad) \; : \; g = A,C \; ; \; \alpha_A = \alpha \; ; \; \alpha_C = 0 \, ,$$

donde ahora la función $f(x)$ pertenece a una clase más general de funciones. Describimos a continuación los modelos de regresión lineal múltiple y los modelos aditivos generalizados (GAM).

4.1. Modelos de regresión lineal múltiple. Supóngase que condicionalmente a cada observación $\mathbf{X}_i = \left(X_{i,1},\cdots,X_{i,p}\right)'$ de un vector p-dimensional se observa una variable aleatoria Y_i, para $i=1,\cdots,n$, El modelo de regresión múltiple supone que $Y_1,\cdots,Y_n$ son variables aleatorias independientes tales que $Y_i \cong N\left(\mu\left(X_{i,1},\cdots,X_{i,p}\right);\sigma\right)$. El modelo es de *regresión lineal múltiple* si $\mu\left(X_{i,1},\cdots,X_{i,p}\right)=\beta_0+\beta_1\cdot X_{i,1}+\cdots+\beta_p\cdot X_{i,p}$. Cabe en este punto repetir que tal aproximación lineal es poco verosímil, aunque tiene el atractivo de ser fácilmente interpretable.

Para el conjunto de datos $\left\{\left(X_{i,1},\cdots,X_{i,p};Y_i\right) : i=1,\cdots,n\right\}$ el parámetro $\theta=\left(\beta_0,\beta_1,\cdots,\beta_p;\sigma^2\right)$ puede estimarse a partir de la función de log-verosimilitud:

$$\ell\left(\beta_0,\beta_1,\cdots,\beta_p;\sigma^2\right)=-\frac{n}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X_{i,1}-\cdots-\beta_pX_{i,p}\right) .$$

La existencia del estimador de máxima verosimilitud $\hat{\theta}_n$ requiere que las variables explicativas $X_{i,1},\cdots,X_{i,p}$ sean linealmente independientes. El problema de la *multicolinealidad* surge cuando entre tales variables exista aproximadamente una relación lineal (una de las variables está altamente correlada con una combinación lineal de las otras). En tal caso, las componentes del estimador $\hat{\theta}_n$ tendrán altas varianzas (pequeñas variaciones de los datos producirán grandes variaciones en las estimaciones). Desde el punto de vista epidemiológico, lo que habitualmente interesa es determinar un conjunto de *variables o factores independientes que expliquen la variable de respuesta* Y_i. No interesa, por tanto, en la práctica forzar directamente la construcción de un modelo $\mu\left(X_{i,1},\cdots,X_{i,p}\right)=\beta_0+\beta_1\cdot X_{i,1}+\cdots+\beta_p\cdot X_{i,p}$ con todas las variables, sino que éstas deben seleccionarse a través de un algoritmo en pasos sucesivos. Los paquetes estadísticos incluyen diversos algoritmos de selección de variables (*forward, backward, stepwise,* etc.). Por su importancia en la construcción de modelos en epidemiología, damos ahora un procedimiento *forward* de construcción de un modelo de regresión lineal.

Considérese, por tanto, que se desea explicar una variable Y a partir de un conjunto de variables explicativas $X_{i,1},\cdots,X_{i,p}$. Para ello se dispone del conjunto de datos $\left\{\left(X_{i,1},\cdots,X_{i,p};Y_i\right): i=1,\cdots,n\right\}$.

1. Para cada variable $X_{i,j}$ se estima el modelo de regresión $E\left[Y_i\middle|X_{i,j}\right] = \beta_0^{(j)}+\beta_1^{(j)}\cdot X_{i,j}$. Se introduce en el modelo aquella variable X_{i,j_0} para la cual el test estadístico F (ANOVA) sea máximo, siempre y cuando el correspondiente p-valor sea significativo.

2. Supóngase que en el modelo se han introducido las variables $X_{i,j_1},\cdots,X_{i,j_r}$ ($E\left[Y_i\middle|X_{i,j_1},\cdots,X_{i,j_r}\right]$). Para cada variable $X_{i,k}$ no introducida en el modelo, se obtienen los F-test (F_k) correspondientes a los modelos $E\left[Y_i\middle|X_{i,j_1},\cdots,X_{i,j_r}\right] = \beta_0+\beta_1X_{i,j_1}+\cdots+\beta_rX_{i,j_r}+\beta_kX_{i,k}$. Se introduce en el modelo aquella variable $X_{i,k}$ para la que el correspondiente F_k sea máximo, siempre y cuando el contraste sea significativo. En caso contrario, se detiene el procedimiento.

Nótese que una variable sólo es introducida en el modelo cuando añade información a las ya contenidas. De este modo, el subconjunto de variables finalmente introducidas constituyen un conjunto de factores que *explican de forma independiente a* (o se asocian de forma independiente con) el *outcome* (Y_i).

Considérese ahora un estudio epidemiológico en el que se desee comparar una determinada exposición sobre una variable Y que supondremos numérica. En la enfermedad diabética, el síndrome metabólico consiste en la alteración de determinados marcadores, entre ellos, los triglicéridos. Analicemos por tanto el efecto de la diabetes sobre los triglicéridos. El siguiente cuadro resume las variables edad y triglicéridos (en escala logarítmica para reducir su asimetría) en los grupos de diabéticos e individuos con tolerancia normal a la glucosa (controles).

	Controles $n = 762$	Diabéticos $n = 129$	p-valor
Edad (años)	45.5 ± 10.8	58.8 ± 10.5	< 0.001
log-*trigliceridos*	4.59 ± 0.48	4.98 ± 0.49	< 0.001

Nótese que en los diabéticos hay un evidente incremento del nivel medio de triglicéridos (8.5%), siendo esta diferencia estadísticamente significativa ($p < 0.001$). Es sabido, por otra parte, que, en general, los niveles lipídicos tienden a incrementarse con el aumento de la edad. Dado que los diabéticos tienen en media 13 años más que los controles, el incremento de triglicéridos en diabéticos podría explicarse total o parcialmente por el hecho de ser la cohorte de diabéticos de mayor edad. Se requeriría, por tanto, hacer la comparación en las mismas condiciones de edad (al menos). Ello es posible en el contexto del modelo de regresión (análisis de la covarianza):

$$E\left[Y\middle|grupo,edad\right] = \theta + \alpha\cdot\left(grupo\right) + \beta\cdot\left(edad\right),$$

donde Y representa la variable log-*triglicéridos* y *grupo* es una variable binaria con los valores 1 (*diabéticos*) y 0 (*controles*). La estimación de este modelo se muestra en la siguiente tabla.

Parámetro	Estimación	Error estándar	p-valor
θ	4.561	0.097	< 0.001
α	0.299	0.049	< 0.001
β	0.007	0.001	< 0.001

A partir de estos resultados pueden obtenerse estimaciones de los valores esperados para cada grupo y edad. Las medias ajustadas por edad se definen habitualmente como el valor esperado para la edad media de todos los sujetos observados en el estudio, en este caso, 47.4 años. Se obtiene de esta forma que la media ajustada para el grupo control es de 4.600 (IC-95% = 4.57; 4.63) y para el grupo de diabéticos de 4.899 (IC-95% = 4.81; 4.99) (nótese que la diferencia de medias es exactamente el parámetro α). Estas medias sí son comparables, al obtenerse en ambos grupos para una misma edad. Nótese que en este caso el incremento del valor medio en el grupo de diabéticos frente al control es de un 6.5% (la variación en las medias no ajustadas fue de un 8.5%), lo que supone que parte de la diferencia es atribuible al hecho de que la edad media en el grupo de diabéticos es superior a la del grupo control.

Nótese que el ajuste se ha realizado mediante el uso de un modelo de regresión lineal. Esto supone que cuando la dependencia se aparta notablemente de la normalidad la diferencia de medias ajustadas (que a la postre marcan si realmente la exposición se asocia con una alteración del marcador) podría ser errónea. Este problema puede resolverse utilizando el ajuste a través de los modelos aditivos generalizados (GAM) que se desarrollan en la siguiente sección.

4.2. Modelos aditivos generalizados. Un modo más flexible de analizar la asociación de un conjunto de variables $X_{i,1},\cdots,X_{i,p}$ sobre un *outcome* Y es a través de los modelos aditivos generalizados (GAM). En el caso de que Y sea una variable numérica, el modelo tiene la forma:

(GAM) $$E\left[Y_i \middle| X_{i,1},\cdots,X_{i,p}\right] = \alpha + s_1\left(X_{i,1}\right)+\cdots+s_p\left(X_{i,p}\right),$$

donde $s_j\left(x\right)$ son funciones no especificadas *suaves* (no paramétricas en el sentido de que se estimarán mediante procedimientos no paramétricos). En el supuesto que Y sea una variable binaria (indicatriz de un evento), el modelo GAM tiene la forma

$$g\left(\mathbb{P}\left(Y_i=1\middle|X_{i,1},\cdots,X_{i,p}\right)\right)=\alpha+s_1\left(X_{i,1}\right)+\cdots+s_p\left(X_{i,p}\right)$$

para alguna función g. El modelo logístico GAM se corresponde con el caso en el que g es la función logit. Los modelos log-lineales de uso frecuente en los problemas de evolución de tasas de mortalidad se obtienen para $g\left(x\right)=\log x$.

Si las funciones $s_j\left(x\right)$ dependieran de un número finito de parámetros, el modelo podría estimarse por mínimos cuadrados o máxima verosimilitud. Parece no obstante más conveniente, al menos en una primera fase, estimar estas funciones de forma no paramétrica, sin perjuicio de imponer sobre ellas condiciones de diferenciabilidad. Esto permitirá detectar de modo más verosímil la forma en la que las variables $X_{i,1},\cdots,X_{i,p}$ se asocian con Y_i.

Consideramos en primer lugar el ajuste para el caso en el que Y_i es una variable numérica. Un criterio de estimación es el de *mínimos cuadrados penalizados*, el cual es

similar al dado en (3.2) para el caso unidimensional. En este caso, la suma de residuales cuadráticos penalizados tiene la forma:

$$PRSS\left(\alpha, s_1, \cdots, s_p\right) = \sum_{i=1}^{n} \left\{ Y_i - \alpha - \sum_{j=1}^{p} s_j\left(X_{i,j}\right) \right\}^2 + \sum_{j=1}^{p} \lambda_j \int \left\{ s_j''\left(t_j\right) \right\}^2 dt_j \ ,$$

donde los λ_j representan los parámetros de suavizamiento. Puede ahora también probarse que la solución del problema:

(4.1) $$\min_{\alpha, s_1, \cdots, s_p} PRSS\left(\alpha, s_1, \cdots, s_p\right)$$

es un spline aditivo cúbico en el sentido de que cada función $s_j\left(x\right)$ es un spline cúbico en la j-ésima componente, con nodos en los puntos $\left\{ X_{i,j} : i = 1, \cdots, n; j = 1, \cdots, p \right\}$. Se requieren, no obstante, condiciones adicionales para la unicidad de la solución. Condiciones habituales son las siguientes:

i. $\sum_{i=1}^{n} s_j\left(X_{i,j}\right) = 0 \ , \ \forall j = 1, \cdots, p$.

ii. La matriz $\left[X_{i,j}\right]_{i,j}$ es no singular (las variables $X_{i,1}, \cdots, X_{i,p}$ son linealmente independientes).

Bajo las condiciones anteriores, la solución de (4.1) es única. Como hecho curioso añadimos que en el caso de que la matriz $\left[X_{i,j}\right]_{i,j}$ sea singular, la solución para la parte lineal de los splines no es única, pero sí lo es la no lineal. Un método para obtener la solución al problema (4.1) (única si se dan las mencionadas condiciones i y ii) es el dado por el siguiente algoritmo (*backfitting*) debido a Hastie y Tibshirani (1990).

Algoritmo backfitting para modelos aditivos

1. Inicializar: $\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} Y_i \ , \ \hat{s}_j \equiv 0, \ \forall j = 1, \cdots, p$.
2. Iterar: $j = 1, 2, \cdots, p, \cdots, 1, 2, \cdots, p, \cdots \qquad \hat{s}_j \equiv \mathbb{S}_j \left(\left\{ Y_i - \hat{\alpha} - \sum_{k \neq j} \hat{s}_k\left(X_{i,k}\right) \right\}^2 \right),$

   donde $\mathbb{S}_j$ representa a cualquiera de los alisadores (núcleo o spline cúbico) definidos en (3.2). En las aplicaciones expuestas en este trabajo se utilizarán los splines. Una vez obtenida la función $s_j\left(x\right)$, ésta se centra en la forma:

$$s_j\left(x\right) \leftarrow s_j\left(x\right) - \frac{1}{n} \sum_{i=1}^{n} s_j\left(X_{i,j}\right).$$

Este proceso se continúa hasta que las funciones varíen $s_j(x)$ por debajo de un umbral especificado.

Para el caso de variables binarias, el criterio de ajuste puede basarse en la función de log-verosimilitudes penalizadas. El algoritmo *backfitting* correspondiente está desarrollado en Hastie y Tibshirani (1990 y 2001).

Considérese nuevamente el estudio de simulación de la evolución de la masa ósea (DXA) en mujeres asmáticas y controles. El siguiente gráfico muestra conjuntamente el patrón de evolución (azul), el ajuste lineal (segmentos en marrón) y el ajuste mediante spline cúbico (rojo).

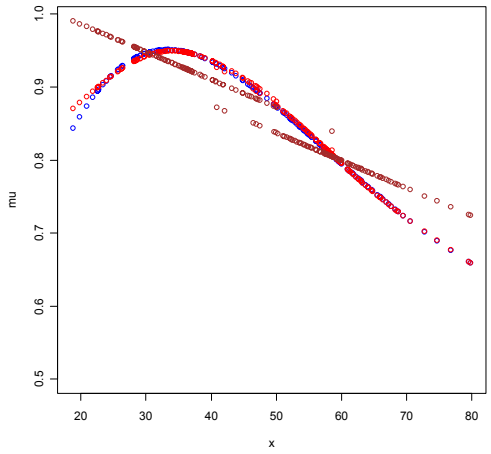

Figura 3. Representación simultánea del patrón de evolución de DXA L2-L4 (simulación), y las estimaciones lineales y no paramétricas.

En el siguiente cuadro se muestran las medias observadas por grupo, las ajustadas a través del modelo lineal y las ajustadas mediante un spline cúbico.

Ajuste	Controles $n = 100$	Casos $n = 100$	p-valor
Ninguno	0.920 (0.004)	0.799 (0.006)	< 0.001 (*)
Lineal	0.883 (0.006)	0.835 (0.006)	< 0.001 (**)
Spline	0.893 (0.004)	0.889 (0.004)	0.512 (***)

Estimación (error estándar); (*) t-test; (**) y (***) F-test

Obsérvese la pobre aproximación que da el modelo lineal al patrón generador de los datos. Además, el efecto de grupo (asmáticas *vs* controles) aparece como estadísticamente significativo cuando se utiliza este ajuste. Por el contrario, el ajuste correspondiente al modelo aditivo (GAM) se adapta perfectamente al patrón original, siendo ahora no significativo el efecto de grupo ($p = 0.512$).

Análisis de los datos SIMU-1. Para los datos SIMU-1 descritos en el segundo epígrafe, las variables Y_i están normalmente generadas, siendo:

$$E\left[Y_i \middle| X_{i,1}, X_{i,2}, X_{i,3}\right] = 4 + 2x_{i,1} + 5\cdot\left(x_{2.i} - 5\right)^2 + 10\cdot\cos\left(3\cdot x_{3,i}\right).$$

Comparamos ahora el ajuste lineal con el aditivo generalizado (GAM). El siguiente cuadro corresponde a la estimación lineal (R).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
------------------------------------------------------------
(Intercept)   2.9814     4.3761   0.681    0.496
x1            2.7046     0.5928   4.563  7.4e-06 ***
x2            0.3200     0.6075   0.527    0.599
x3            0.3075     0.2746   1.120    0.264
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

Nótese que la única estimación aceptable es la que corresponde a la parte lineal del patrón generador. Los efectos de las variables $X_{i,2}$ y $X_{i,3}$ no son detectados ya que (sobre todo el de $X_{i,3}$) se apartan de la linealidad. La Figura 4.a muestra para cada observación Y_i la correspondiente predicción del modelo. La simple observación de esta gráfica permite calificar el ajuste como desastroso. El siguiente cuadro muestra la salida de resultados del paquete R correspondiente al ajuste por un modelo aditivo generalizado (GAM) utilizando el algoritmo *backfitting*. Los p-valores se determinan mediante test de razón de verosimilitudes. La abreviación *edf* representa grados de libertad estimados (ver Hastie y Tibshirani, 1990, apéndice B). En la Figura 4.b puede apreciarse la evidente ventaja de los GAM sobre el ajuste lineal.

```
Parametric coefficients:
              Estimate  std. err.    t ratio    Pr(>|t|)
------------------------------------------------------------
(Intercept)     19.641    0.06771      290.1    < 2.22e-16

Approximate significance of smooth terms:
             edf       chi.sq     p-value
s(x1)      1.543       903.28     < 2.22e-16
s(x2)      7.331        12341     < 2.22e-16
s(x3)      8.998        10828     < 2.22e-16
```

La Figura 5.a sugiere que el efecto de la primera variable $X_{i,1}$ es lineal. El epidemiólogo puede estar interesado en confirmar este extremo. Un modo de hacerlo es ajustar el modelo con la restricción de que este efecto es lineal y, utilizando los residuales, contrastar si existe un efecto no lineal. El siguiente cuadro muestra el ajuste bajo la restricción de linealidad para $X_{i,1}$.

```
Parametric coefficients:
             Estimate  std. err.    t ratio    Pr(>|t|)
-------------------------------------------------------------
(Intercept)    9.5074     0.3447      27.58    < 2.22e-16
         x1    2.0236    0.06749      29.98    < 2.22e-16

Approximate significance of smooth terms:
             edf       chi.sq     p-value
s(x2)      7.348        12304     < 2.22e-16
s(x3)      8.998        10807     < 2.22e-16
```

Nótese que la estimación del coeficiente de $X_{i,1}$ coincide de un modo muy preciso con el valor real. El siguiente cuadro muestra el ajuste de los residuales $Y_i - \hat{\mu}_i$ frente a un $X_{i,1}$ a través de un spline cúbico en orden a determinar si existe un efecto no lineal de esta variable sobre la respuesta.

```
Parametric coefficients:
             Estimate  std. err.    t ratio    Pr(>|t|)
(Intercept) 2.0358e-13    0.06583 3.093e-12    1

Approximate significance of smooth terms:
             edf       chi.sq     p-value
s(x1)      1.545      0.97147     0.49449
```

La no significación estadística ($p = 0.49449$) permite confirmar que la contribución de esta variable sobre el *outcome* es lineal.

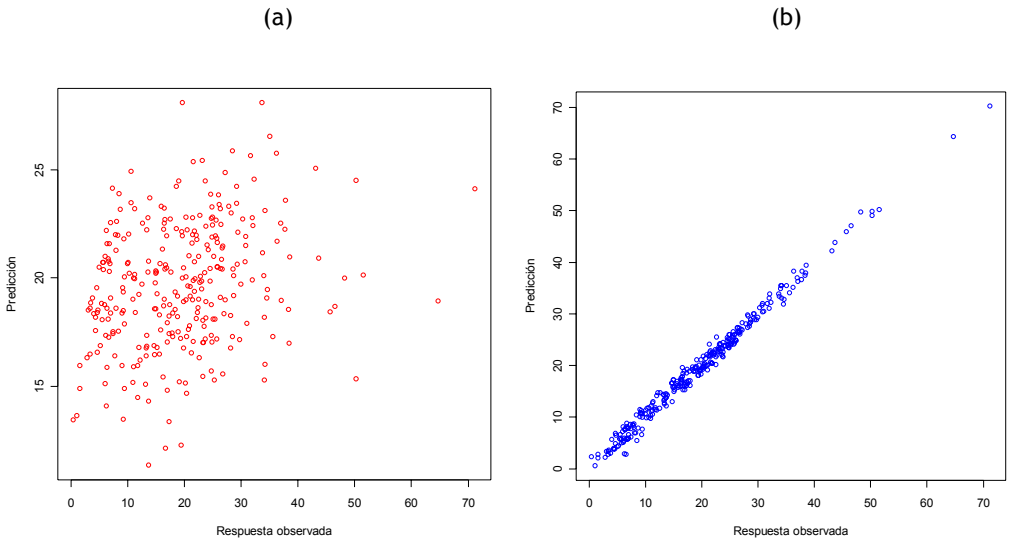

Figura 4. Predicciones del modelo frente a los valores observados. (a) Ajuste lineal; (b) ajuste GAM.

La Figura 5 muestra la estimación mediante splines cúbicos de las tres componentes correspondientes al estudio de simulación SIMU-1.

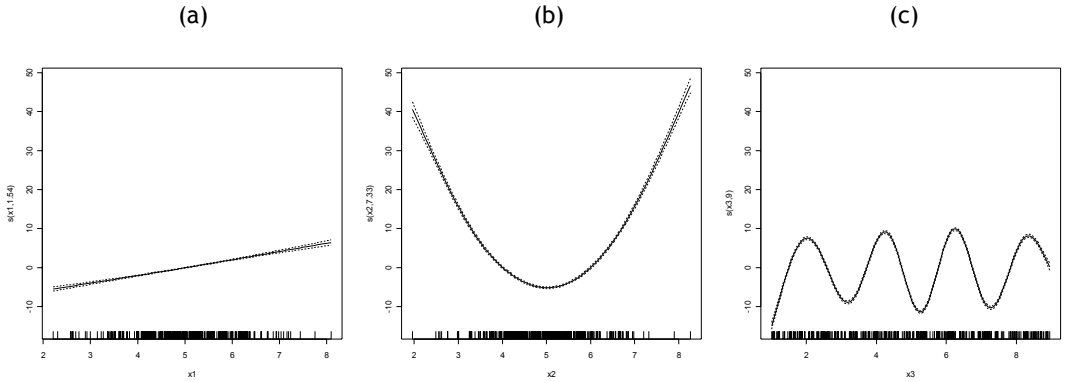

Figura 5. Estimación de las componentes de los datos de SIMU-1.

Estudio de diabetes. Analizamos ahora la asociación de la diabetes mellitus tipo 2 (DM2) con el siguiente un conjunto de factores: *cint* (longitud de la cintura), *trig* (triglicéridos), *antd* (variable binaria indicatriz de antecedentes diabéticos), *sexpcr* (interacción sexo*PCR; valor 1 en varones con proteína reactiva C (PCR) elevada y 0 en el resto) y *edad*. $DM2$ representa la variable con valores 1 (***diabetes***) y 0 y (***tolerancia normal a la glucosa***). Exploraciones previas de los datos sugieren ajustar algunos factores de forma lineal (cintura y triglicéridos). El modelo propuesto es:

$$\mathbb{P}\left(DM2=1 \middle| \, cint,trig,antd,sexpcr,edad\right) =$$
$$\beta_0+\beta_1\cdot cint+\beta_2\cdot trig+\beta_3\cdot antd+\beta_4\cdot sexpcr+s_e\left(edad\right)+s_t\left(trig*sexo\right);$$

*trig*sexo* representa la interacción entre el nivel de triglicéridos y sexo. Esta variable coincide con los triglicéridos sólo en la cohorte de hombres, siendo nula en la de mujeres. El siguiente cuadro muestra la estimación del modelo.

```
 Parametric coefficients:
              Estimate  std. err.   t ratio    Pr(>|t|)      OR IC-95%
-------------------------------------------------------------------------------
(Intercept)    -10.098       1.16    -8.708    < 2.22e-16
       cint   0.057512    0.01048     5.488    4.0746e-08  1.059 (1.038;1.081)
       trig  0.0043901    0.00213     2.061    0.039305    1.004 (1.000;1.009)
       antd     1.0318      0.244     4.229    2.3465e-05  2.806 (1.739;4.527)
     sexpcr     1.4248      0.578     2.465    0.013698    4.157 (1.339;12.91)

Approximate significance of smooth terms:
                   edf       chi.sq     p-value
     s(edad)     3.103       79.699     < 2.22e-16
s(trig):sexo         1       4.1245     0.042268
```

Las *odd-ratios* se han estimado sólo para los efectos lineales (en las variables numéricas expresan el aumento de riesgo por incremento de la variable explicativa). Los efectos no lineales se muestran en la Figura 6.a. Puede observarse que a partir de los 60 años se acelera el riesgo de la enfermedad y que éste se estabiliza a partir de los 65 años. La interacción entre sexo y triglicéridos es significativa ($p = 0.042268$) y la Figura 6.b indica que en la mujer los triglicéridos tienen una mayor asociación con la DM2, confirmándose de esta forma la interacción *trig*sexo*.

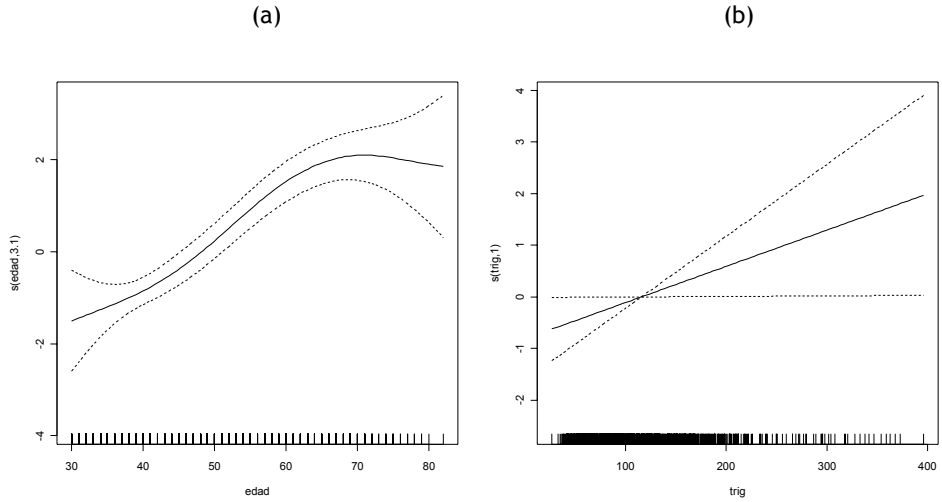

Figura 6. Estimación de los efectos de la edad (a) y la interacción de triglicéridos con el sexo (b).

Podemos concluir que los antecedentes diabéticos se asocian con la DM2 mientras que la elevación de la PCR se asocia sólo en hombres. Por cada unidad de aumento de la longitud de cintura, se incrementa el riesgo de diabetes en un grado dado por una OR = 1.059 (IC-95% = 1.038; 1.081). En hombres, el aumento de riesgo por unidad de incremento en los triglicéridos viene dado por una OR = 1.004 (IC-95% = 1; 1.009). El incremento de los triglicéridos en la mujer conduce a un mayor aumento de riesgo de diabetes ($p = 0.042268$). El aumento de riesgo a lo largo de las edades no puede darse a través de una OR, dado que éste no se produce de forma lineal (se acelera alrededor de los 40 años y se detiene en los 70).

Referencias

M. Boronat, V.F. Varillas, P. Saavedra, V. Suárez, E. Bosch, A. Carrillo, F.J. Novoa: Diabetes mellitus and impaired glucose regulation in the Canary Islands (Spain): prevalence and associated factors in the adult population of Telde, Gran Canaria. *Diabetic Medicine* (2005) [aceptado].

M.J. Campbell, D. Machin: *Medical statistics* (2nd ed.). Wiley, Chichester, 1993.

W. Härdle: *Smoothing techniques. With implementation in S*. Springer-Verlag, New York, 1991.

T.J. Hastie, R.J. Tibshirani: *Generalized additive models*. Chapman and Hall, New York, 1990.

T.J. Hastie, R.J. Tibshirani, J. Friedman: *The elements of statistical learning: Data, inference and prediction*. Springer-Verlag, New York, 2001.

D.W. Hosmer, S. Lemeshow: *Applied logistic regression* (2nd ed.). Wiley, New York, 2000.

D. Peña Sánchez de Rivera: *Estadística: modelos y métodos*. Alianza Universidad Textos, Madrid, 1987.

J. del Rey Calero: *Método epidemiológico y salud de la comunidad*. Interamericana McGraw-Hill, Madrid, 1989.

S. Word, N. Augustin: GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Preprint* (2002).

Apéndice

En este apéndice se muestran detalles matemáticos de interés para la modelación de conjuntos de datos independientes de la forma $\left\{\left(X_{i,1},\cdots,X_{i,p};Y_i\right) : i=1,\cdots,n\right\}$. Para el modelo de regresión lineal múltiple mostramos en *A.1* que el estimador del vector de medias $\boldsymbol{\mu}$ es la proyección del vector de observaciones $\mathbf{Y}$ sobre el espacio vectorial generado por los vectores columna $\mathbf{x}_j'=\left(X_{1,j},\cdots,X_{n,j}\right)$ (*j*-ésima variable explicativa evaluada a lo largo de los objetos). En el epígrafe *A.2* se considera el caso en el que $p=1$ (un solo predictor), pero en el que la respuesta se modela a través de un conjunto de funciones base especificadas $b_1(X),\cdots,b_p(X)$. El problema de la estimación es similar al expuesto en *A.1* en el sentido de que el estimador del vector de medias $\boldsymbol{\mu}$ es ahora la proyección sobre el espacio vectorial generado por los vectores $\left(b_j(X_1),\cdots,b_j(X_n)\right)'$, para $j=1,\cdots,p$. Un caso especial de estimación mediante funciones base es el de los splines naturales cúbicos. En la práctica, el uso de este procedimiento de ajuste plantea el problema de la elección de nodos. Curiosamente, el procedimiento de los mínimos cuadrados penalizados expuesto en el epígrafe *A.3* tiene como solución un spline natural cúbico cuyos nodos coinciden con los puntos del diseño $X_i : i=1,\cdots,n$. El problema de los nodos se traslada entonces a la selección del parámetro de alisamiento. Este problema es tratado en *A.4* a través del método de validación cruzada. La complejidad del algoritmo *backfitting* utilizado para la estimación de los modelos aditivos se documenta en *A.5*. Finalmente, en *A.6* se consideran algunos contrastes de interés en el contexto de los modelos aditivos.

A.1. Estimación del modelo de regresión lineal múltiple

Considérese una muestra aleatoria de n objetos de una población y que para un conjunto de variables $X_{i,1},\cdots,X_{i,p} : \ i=1,\cdots,n$ se observa una variable aleatoria Y_i con ley de

probabilidad $N(\mu_i,\sigma)$, donde $\mu_i=\beta_0+\beta_1X_{i,1}+\cdots+\beta_pX_{i,p}$ y de tal forma que $Y_1,\cdots,Y_n$ son independientes. En orden a la manipulación algebraica del modelo, conviene utilizar la siguiente notación matricial:

$$\mathbf{Y}=\begin{pmatrix}Y_1\\Y_2\\...\\Y_n\end{pmatrix}\qquad \mathbf{X}=\begin{pmatrix}1 & X_{1,1}\cdots & X_{1,p}\\1 & X_{2,1}\cdots & X_{2,p}\\ & \cdots\cdots & \\1 & X_{n,1}\cdots & X_{n,p}\end{pmatrix}\quad \boldsymbol{\beta}=\begin{pmatrix}\beta_0\\\beta_1\\\cdots\\\beta_p\end{pmatrix}\qquad \mathbf{I}=\begin{pmatrix}1 & 0 & \cdots & 0\\0 & 1 & \cdots & 0\\ & \cdots\cdots & & \\0 & 0 & \cdots & 1\end{pmatrix}$$

Aquí, la matriz $\mathbf{X}$ recibe el nombre de *matriz de diseño*. El modelo puede entonces expresarse en la forma: $\mathbf{Y}\cong N_n\left(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}\right)$.

Sea $\mathbf{x}_j'=\left(X_{1,j},\cdots,X_{n,j}\right)$ la *j*-ésima columna de la matriz de diseño (correspondiente en realidad a la *j*-ésima variable explicativa observada) y sea además $\mathbf{u}'=(1,\cdots,1)$. La existencia del estimador máximo verosímil para el parámetro $\boldsymbol{\beta}$ requiere que, en el espacio $\mathbb{R}^n$, los vectores $\mathbf{u},\mathbf{x}_1,\cdots,\mathbf{x}_p$ sean linealmente independientes. En tal caso, el estimador de máxima verosimilitud (ó mínimos cuadrados) se obtiene como solución del problema $\min_{\boldsymbol{\beta}}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|$. La forma del estimador máximo verosímil es:

$$\hat{\boldsymbol{\beta}}_n=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y},$$

de donde el estimador de $\boldsymbol{\mu}'=\left(\mu_1,\cdots,\mu_p\right)$ es $\boldsymbol{\mu}_n=\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}=\mathbf{S}\mathbf{Y}$. Nótese que la condición de independencia lineal de $\mathbf{u},\mathbf{x}_1,\cdots,\mathbf{x}_p$ garantiza la invertibilidad de la matriz $\mathbf{X}'\mathbf{X}$, y con ello la existencia de un único estimador máximo verosímil para el parámetro $\boldsymbol{\beta}$. Puede también comprobarse que la matriz $\mathbf{S}$ corresponde a la matriz proyección sobre el subespacio vectorial definido por los vectores linealmente independientes $\mathbf{u},\mathbf{x}_1,\cdots,\mathbf{x}_p$. Es fácil probar que $\hat{\boldsymbol{\beta}}_n\cong N\left(\boldsymbol{\beta},\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$ y de aquí, $\hat{\boldsymbol{\mu}}_n\cong N_n\left(\boldsymbol{\mu},\sigma^2\mathbf{S}\right)$ donde $\mathbf{S}=\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ es la matriz proyección (*sombrero*). Un estimador centrado para la varianza σ^2 es:

$$\hat{\sigma}_n^2=\frac{1}{n-\mathrm{tr}(\mathbf{S})}\|\mathbf{Y}-\hat{\boldsymbol{\mu}}\|^2,$$

donde $\mathrm{tr}(\mathbf{S})$ representa la traza de la matriz proyección $\mathbf{S}$. Veamos que el estimador es centrado:

$$E\left[\hat{\sigma}_n^2\right]=\frac{1}{n-\mathrm{tr}(\mathbf{S})}E\left[\left\|\mathbf{Y}-\hat{\boldsymbol{\mu}}\right\|^2\right]=\frac{1}{n-\mathrm{tr}(\mathbf{S})}E\left[\left\|(\mathbf{I}-\mathbf{S})\mathbf{Y}\right\|^2\right]=\frac{\mathrm{tr}(\mathbf{I}-\mathbf{S})}{n-\mathrm{tr}(\mathbf{S})}\sigma^2=\sigma^2.$$

En general, $\mathrm{tr}(\mathbf{S})$ se corresponde con la dimensión del espacio generado por la familia de vectores linealmente independientes $\mathbf{u},\mathbf{x}_1,\cdots,\mathbf{x}_p$ y, de esta forma, es $p+1$.

A.2. Modelación mediante funciones base: splines cúbicos

Considérese ahora un conjunto de datos $\{(X_i,Y_i):i=1,\cdots,n\}$ de tal forma que, condicionalmente a X_i, la variable aleatoria Y_i satisface $Y_i\cong N(\mu(X_i),\sigma)$, siendo:

$$\mu(x)=\sum_{j=1}^{p}\alpha_j b_j(x),$$

donde las funciones *base* $b_j(x):\ j=1,\cdots,p$ están plenamente especificadas. El problema de ajuste es idéntico al dado en *A.1*, pero siendo ahora la matriz de diseño:

$$\mathbf{B}=\begin{pmatrix} b_1(X_1) & b_2(X_1) & \cdots & b_p(X_1)\\ b_1(X_2) & b_2(X_2) & \cdots & b_p(X_2)\\ \cdots & \cdots & \cdots & \cdots\\ b_1(X_n) & b_2(X_n) & \cdots & b_p(X_n)\end{pmatrix}.$$

En la sección 3.2 se definieron las bases de splines cúbicos y se mostró la base elemental determinada por un conjunto de nodos $\xi_1<\cdots<\xi_K$, siendo $b_j(x)=\left|x-\xi_j\right|^3$ para $j=1,\cdots,K$, $b_{K+1}(x)=1$ y $b_{K+2}(x)=x$. En general, el ajuste mediante polinomios tiende a ser errático en las proximidades de las fronteras (explosión de la varianza del estimador), y obviamente este problema se complica con los splines cúbicos. Una forma de resolverlo es forzando a que las derivadas segundas de $\mu(x)$ se anulen fuera del intervalo $[\xi_1,\xi_K]$. En tal caso, el spline se denomina *spline natural cúbico*. Para la base considerada, una función de la forma:

$$\mu(x)=\sum_{j=1}^{K+2}\alpha_j b_j(x),$$

donde los coeficientes están sujetos a las restricciones $\sum_{j=1}^{K}\alpha_j = 0$ y $\sum_{j=1}^{K}\alpha_j\xi_j = 0$, es un spline natural cúbico. Tales restricciones evidentemente rebajan el número efectivo de parámetros a K.

A.3. Mínimos cuadrados penalizados

Para la suma de residuales cuadráticos penalizados:

$$\mathrm{PRSS}(\mu,\lambda) = \sum_{i=1}^{n}\left\{ Y_i - \mu(x_i) \right\}^2 + \lambda\int\left\{\mu''(x)\right\}^2 dx\,,$$

puede probarse que la solución del problema $\min_{\mu}\mathrm{PRSS}(\mu,\lambda)$ tiene una solución explícita, finito dimensional, la cual es un spline natural cúbico con nodos únicos en el conjunto de valores $\{X_i : i = 1,\cdots,n\}$. Si $\mathbf{B}$ es la matriz cuyas columnas son las funciones base calculadas a lo largo de los valores X_i, la solución del problema tiene la forma $\hat{\boldsymbol{\mu}} = \mathbf{B}\left(\mathbf{B}'\mathbf{B} + \lambda\boldsymbol{\Omega}\right)^{-1}\mathbf{B}'\mathbf{Y}$, siendo $\boldsymbol{\Omega} = \left(\int b_j''(x)b_k''(x)dx\right)_{j,k}$.

Aparentemente, el hecho de que el número de nodos coincida con el de datos conduce a una sobre-parametrización del modelo. Sin embargo, la penalización impuesta reduce el número efectivo de parámetros. En el ajuste por un modelo de regresión lineal múltiple, la traza de la matriz proyección $\mathbf{S} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ coincide con el número de parámetros del modelo. En general, en un problema de ajuste en el que el estimador es de la forma $\hat{\boldsymbol{\mu}} = \mathbf{S}\mathbf{Y}$, el grado de alisamiento efectivo (*grados de libertad efectivos*) se define como la traza de la matriz $\mathbf{S}$. Por tanto, el estimador de mínimos cuadrados penalizado viene dado por $\mathrm{tr}\left(\mathbf{B}\left(\mathbf{B}'\mathbf{B} + \lambda\boldsymbol{\Omega}\right)^{-1}\mathbf{B}'\right)$.

A.4. Selección del parámetro de alisamiento por validación cruzada generalizada.

El problema de la selección de nodos en el ajuste por splines cúbicos se resolvió mediante el planteamiento alternativo del problema de ajuste utilizando los residuales cuadráticos penalizados. La solución de este problema coincide con un spline cúbico con un único conjunto de nodos formados por los puntos del diseño $X_1,\cdots,X_n$. Como cabía esperar, el problema de la elección de nodos se traslada al problema de la selección del parámetro de alisamiento λ. Por tal motivo, de ahora en adelante expresamos la dependencia del estimador del parámetro de alisamiento por $\hat{\mu}_\lambda(x)$.

Un criterio natural de selección consiste en determinarlo de tal forma que el *promedio del error cuadrático medio* sea mínimo. Sea por tanto:

$$\mathrm{MSE}(\lambda)=\frac{1}{n}\sum_{i=1}^{n}E\left[\left(\hat{\mu}_\lambda(x)-\mu(x)\right)^2\right].$$

De aquí, consideraremos como parámetro de suavizamiento óptimo el valor λ_0 tal que $\mathrm{MSE}(\lambda_0)=\min_\lambda \mathrm{MSE}(\lambda)$. En orden a su estimación consideramos en primer lugar el llamada *promedio error cuadrático predictivo*, definido por

$$\mathrm{PSE}(\lambda)=\frac{1}{n}\sum_{i=1}^{n}E\left[\left(Y_i^*-\hat{\mu}_\lambda(x_i)\right)^2\right],$$

donde Y_i^* es una nueva observación en x_i (diferente de las que se utilizaron para la estimación $\hat{\mu}_\lambda(x)$. Es fácil probar que $\mathrm{PSE}(\lambda)=\mathrm{MSE}(\lambda)+\sigma^2$. De esta forma, el valor de λ que minimiza $PSE(\lambda)$ coincide con λ_0. Por otra parte, $PSE(\lambda)$ puede aproximarse por la función:

$$\mathrm{CV}(\lambda)=\frac{1}{n}\sum_{i=1}^{n}\left\{\ Y_i-\hat{\mu}_\lambda^{-i}(x_i)\right\}^2 .$$

Esta función recibe el nombre de *función de validación cruzada*, y la expresión $\hat{\mu}_\lambda^{-i}(x_i)$ se refiere al estimador de $\mu(x)$ obtenida sin el punto (x_i,Y_i). A partir de esta función puede estimarse λ_0 como el valor $\hat{\lambda}$ tal que $\mathrm{CV}\left(\hat{\lambda}\right)=\min_\lambda \mathrm{CV}(\lambda)$. El fundamento de esta estimación se basa en la aproximación $E\left[\mathrm{CV}(\lambda)\right]\approx \mathrm{PSE}(\lambda)$. Puede probarse además:

$$\mathrm{CV}(\lambda)=\frac{1}{n}\sum_{i=1}^{n}\left\{\ \frac{Y_i-\hat{\mu}_\lambda(x_i)}{1-S_{i,i}}\right\}^2 ,$$

siendo $S_{i,i}$ el i-ésimo elemento de la diagonal de la matriz proyección $\mathbf{S}$. Alternativamente a la función de validación cruzada, se define la *función de validación cruzada generalizada* por:

$$\mathrm{GCV}(\lambda)=\frac{1}{n}\sum_{i=1}^{n}\left\{\ \frac{Y_i-\hat{\mu}_\lambda(x_i)}{1-\mathrm{tr}(\mathbf{S})/n}\right\}^2 .$$

En algunas situaciones puede obtenerse más fácilmente $\mathrm{tr}(\mathbf{S})/n$ que los valores $S_{i,i}$ individuales.

A.5. Modelos aditivos: algoritmo backfitting

Para el ajuste del conjunto de datos $\left\{\left(X_{i,1},\cdots X_{i,p};Y_i\right): i=1,\cdots,n\right\}$ al modelo aditivo $E\left[Y_i\middle|X_{i,1},\cdots,X_{i,p}\right]=f_1\left(X_{i,1}\right)+\cdots+f_p\left(X_{i,p}\right)$ puede determinarse mediante la resolución del problema

$$\begin{pmatrix} \mathbf{I} & \mathbf{S_1} & \mathbf{S}_1 & \cdots & \mathbf{S_1} \\ \mathbf{S_2} & \mathbf{I} & \mathbf{S_2} & \cdots & \mathbf{S_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_p & \mathbf{S}_p & \mathbf{S}_p & \cdots & \mathbf{I} \end{pmatrix}\begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_p \end{pmatrix}=\begin{pmatrix} \mathbf{S}_1\mathbf{Y} \\ \mathbf{S}_2\mathbf{Y} \\ \vdots \\ \mathbf{S}_p\mathbf{Y} \end{pmatrix}$$

donde $\mathbf{S}_j$ es la matriz de alisamiento correspondiente a la *j*-ésima función (ver Hastie and Tibshirani, 1990). Este sistema pude resolverse utilizando, por ejemplo, la descomposición QR. En tal caso, dado que cada matriz $\mathbf{S}_j$ es de orden *n*, el sistema es de orden $np\times np$, lo cual significa que su resolución requiere $O\left((np)^3\right)$ operaciones.

A.6. Contrastes de significación

Para el conjunto de datos $\left\{\left(X_{i,1},\cdots X_{i,p};Y_i\right): i=1,\cdots,n\right\}$ consideremos el modelo $Y_i=\sum_{j=1}^{p} f_j\left(X_{i,j}\right)+e_i$, donde $e_1,\cdots,e_n$ son variables aleatorias independientes con $E[e_i]=0$ y $\mathrm{var}(e_i)=\sigma^2$. En este epígrafe tratamos los problemas de estimación de σ^2, así como de los contrastes $H_{0,j}: f_j\equiv 0$. Una vez obtenidos los estimadores $\hat{f}_j$, consideramos las predicciones $\hat{\mu}_i=\sum_{j=1}^{p}\hat{f}_j\left(X_{i,j}\right)$. En notación vectorial, las predicciones pueden expresarse por $\hat{\boldsymbol{\mu}}=\mathbf{SY}$, donde $\mathbf{S}$ es la matriz de alisamiento. Definimos entonces el estimador:

$$\hat{\sigma}^2=\frac{1}{n-g}\sum_{i=1}^{n}\left(Y_i-\hat{\mu}_i\right)^2,$$

donde $g=\mathrm{tr}(\mathbf{S})$ son los grados de libertad del alisamiento. Alternativamente, puede considerarse $g=\mathrm{tr}\left(2\mathbf{S}-\mathbf{SS}'\right)$ (ver Hastie y Tibshirani, 1990, p. 54). En orden a estudiar las propiedades del estimador $\hat{\sigma}^2$ consideramos las siguientes expresiones:

$$\mathbf{Y}-\hat{\boldsymbol{\mu}}=(\mathbf{I}-\mathbf{S})\mathbf{Y}=(\mathbf{I}-\mathbf{S})(\boldsymbol{\mu}+\mathbf{e})=(\boldsymbol{\mu}-\mathbf{S}\boldsymbol{\mu})+(\mathbf{I}-\mathbf{S})\mathbf{e}\,.$$

Si admitimos que $\hat{\boldsymbol{\mu}}=\mathbf{S}\mathbf{Y}$ es centrado (lo es asintóticamente), se tiene obviamente que $\boldsymbol{\mu}=\mathbf{S}\boldsymbol{\mu}$, y de aquí que $\mathbf{Y}-\hat{\boldsymbol{\mu}}=(\mathbf{I}-\mathbf{S})\mathbf{e}$. De esta forma

$$\hat{\sigma}^2=\frac{1}{n-g}(\mathbf{Y}-\hat{\boldsymbol{\mu}})'(\mathbf{Y}-\hat{\boldsymbol{\mu}})=\frac{1}{n-g}\mathbf{e}'(\mathbf{I}-\mathbf{S})'(\mathbf{I}-\mathbf{S})\mathbf{e}\,,$$

donde $\mathbf{I}$ es la matriz identidad. Así se tiene:

$$E\left[\hat{\sigma}^2\right]=\frac{1}{n-g}E\left[\mathbf{e}'(\mathbf{I}-\mathbf{S})'(\mathbf{I}-\mathbf{S})\mathbf{e}\right]=\frac{1}{n-g}\left(n-\mathrm{tr}(\mathbf{S})\right)\sigma^2=\sigma^2\,.$$

Si admitimos además que $Y_1,\cdots,Y_n$ están normalmente distribuidas tendremos entonces que $(n-g)\hat{\sigma}^2/\sigma^2\cong\chi^2(n-g)$. Puede verse una mejor aproximación a la distribución de probabilidad de este estadístico (nótese que $\hat{\sigma}^2$ no es centrado) también en Hastie y Tibshirani (1990, p. 66).

El problema de contrastar $H_{0,j}:f_j\equiv 0$ puede considerarse un caso especial de comparar dos alisamientos $\hat{\boldsymbol{\mu}}_1=\mathbf{S}_1\mathbf{Y}$ y $\hat{\boldsymbol{\mu}}_2=\mathbf{S}_2\mathbf{Y}$. El segundo modelo podría ser de mayor complejidad que el primero, incluyendo por ejemplo un predictor X_j (y su correspondiente f_j) que no aparece en aquél. Sea $RSS_i=(\mathbf{Y}-\hat{\boldsymbol{\mu}}_i)'(\mathbf{Y}-\hat{\boldsymbol{\mu}}_i)$, para $i=1,2$ y el test estadístico:

$$F=\frac{(RSS_1-RSS_2)/(g_2-g_1)}{RSS_2/(n-g_2)}.$$

Nótese que $RSS_2/(n-g_2)$ es la estimación anteriormente considerada de σ^2 en el contexto del ajuste mediante $\hat{\boldsymbol{\mu}}_2$. Bajo la hipótesis de no significación de los elementos del segundo modelo que no aparecen en el primero, $F\cong f(g_2-g_1,n-g_2)$.

Este tipo de contrastes es especialmente interesante en orden a determinar si una determinada variable influye de forma lineal en el *outcome*. En el contexto de ajustes mediante splines naturales cúbicos, el efecto de una variable X_j se modela en la forma $\beta_j X_j+\sum_j\gamma_j\left|X_j-\xi_j\right|^3$. El contraste $H_{0,j}:\gamma_j=0$, $\forall j$ puede resolverse mediante este tipo de test.