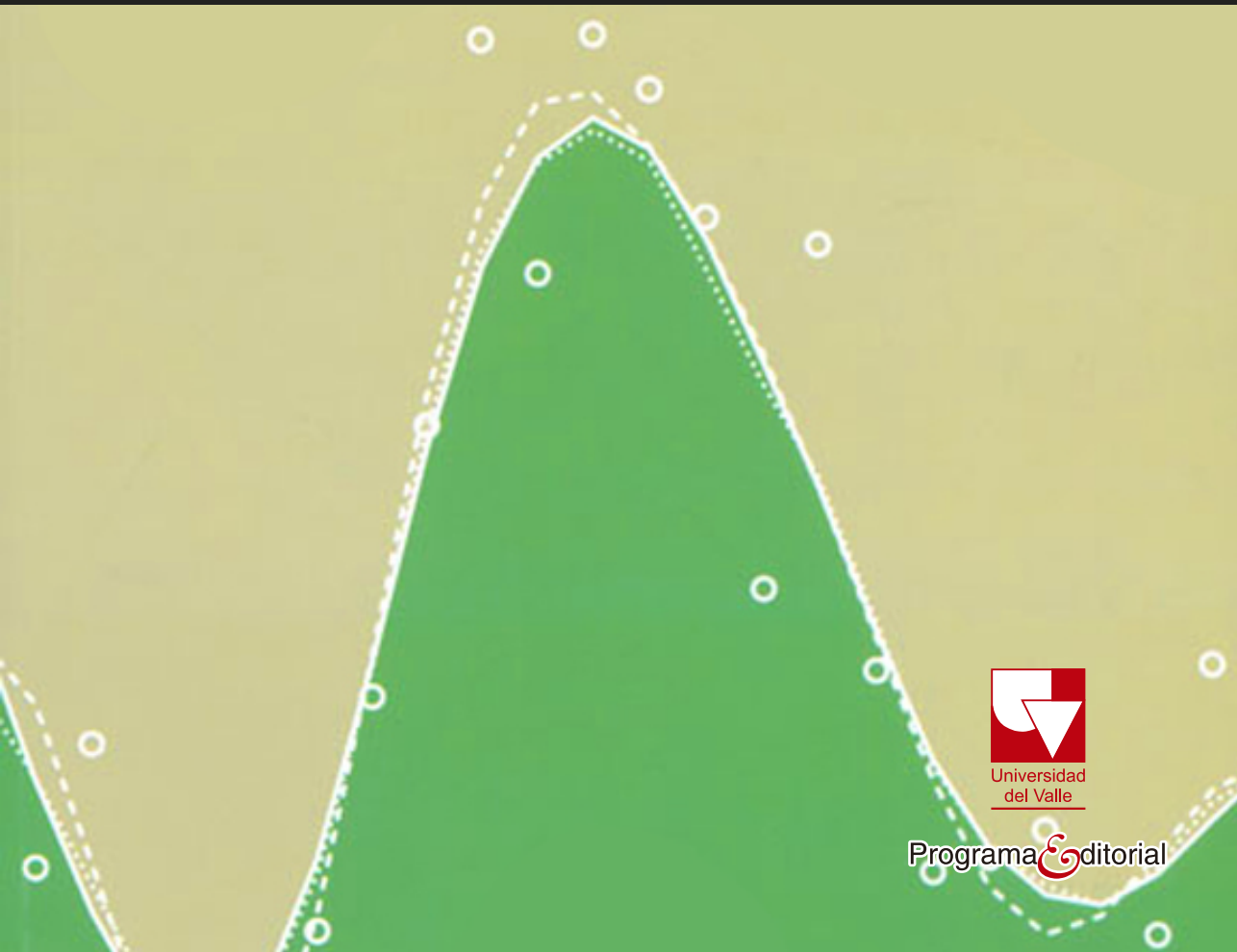
MÉTODOS DE REGRESIÓN NO PARAMÉTRICA

• Javier Olaya Ochoa •



Universidad del Valle
Programa Editorial

La esencia del texto está en intentar comunicar de la manera más amigable posible, sin perder rigurosidad, un conjunto de técnicas que amplían las ideas generales del análisis de regresión. En cuanto a la rigurosidad, los lectores deben estar preparados en ideas básicas de cálculo, álgebra lineal, inferencia estadística paramétrica y no paramétrica y modelos lineales.

Este texto se concentra en el problema de la regresión no paramétrica vista desde diferentes ángulos. El lector encontrará aquí un resumen de los métodos de regresión no paramétrica que se usan con mayor frecuencia, incluyendo la regresión kernel, la suavización spline y la regresión lineal local, siguiendo todos a una introducción a los estimadores de series que presenta las ideas centrales de los métodos.

Universidad del Valle

Programa Editorial

JAVIER OLAYA OCHOA

Colombiano nacido en Tuluá a mediados de los años 1950. Llegó a la Universidad del Valle en los albores de los años 1970 como estudiante del antiguo programa de Laboratorio Químico, del cual obtuvo su título en 1977; se vinculó ese mismo año como funcionario de la Universidad e inició sus estudios de pregrado en Química. Pero cerca de finalizar sus estudios en esta disciplina, decidió iniciar sus estudios de pregrado en el nuevo programa de Estadística, a finales de la misma década. Sus estudios de posgrado en la Universidad de Clemson (EE. UU.) se desarrollaron en el departamento de Matemáticas, en el cual finalizó su Maestría en Ciencias Matemáticas en 1997 y su Doctorado en Management Science en 2000. Actualmente es Profesor Titular de la Escuela de Estadística de la Facultad de Ingeniería y sus principales intereses de investigación son los modelos de regresión, el control estadístico de la calidad, los métodos multivariados, la estadística ambiental y el análisis de encuestas por muestreo.

Métodos de Regresión No Paramétrica

Colección Ingeniería

PÁGINA EN BLANCO
EN LA EDICIÓN IMPRESA

JAVIEROLAYAOCHOA

Métodos de Regresión No Paramétrica

Colección Ingeniería

Olaya Ochoa, Javier
Métodos de regresión no paramétrica / Javier Olaya Ochoa. – Santiago de Cali: Programa Editorial Universidad del Valle, 2012.
124 p. ; 24 cm. – (Ciencias naturales y exactas)
Incluye bibliografía.
1. Modelos lineales (Estadística) 2. Métodos no paramétricos 3. Estadística no paramétrica 4. Análisis de regresión I. Tít.
II. Serie.
519.5 cd 21 ed.
A1378644

CEP-Banco de la República-Biblioteca Luis Ángel Arango

Universidad del Valle
Programa Editorial
Título: Métodos de regresión no paramétrica
Autor: Javier Olaya Ochoa
ISBN: 978-958-765-041-9
ISBN-PDF: 978-958-5164-30-7
DOI: 10.25100/peu.505
Colección: Ingeniería
Primera Edición Impresa diciembre 2012

Rector de la Universidad del Valle: Édgar Varela Barrios
Vicerrector de Investigaciones: Héctor Cadavid Ramírez
Director del Programa Editorial: Omar J. Díaz Saldaña

Diseño de carátula: Anna Echavarria. Elefante
Diagramación: Juan Carlos Pérez Méndez

Cali, Colombia, diciembre de 2020

A Ruby, a mis hijos y a mis mayores.
Mi vida no sería vida sin ellos.

A León Sanjuán,
de quien me gustaría heredar el placer de escribir por gusto.

PÁGINA EN BLANCO
EN LA EDICIÓN IMPRESA

CONTENIDO

ÍNDICE DE TABLAS

PÁGINA EN BLANCO
EN LA EDICIÓN IMPRESA

ÍNDICE DE FIGURAS

PREFACIO

Simon Singh en su libro El Enigma de Fermat (Singh 1997) cita que según Alfred Adler: "La vida matemática de un matemático es corta. Raramente se progresa más allá de los veinticinco años. Si poco se ha logrado hasta entonces, poco se logrará jamás". Y cita a G. H. Hardy, quien según Singh escribió en su libro *A Mathematician's Apology* que: "Los jóvenes demuestran los teoremas, los ancianos escriben los libros".

Yo no soy ni matemático ni joven y el único teorema que enuncié y probé jamás ha sido citado. Y no estoy seguro si clasifico como anciano, pero sí escribí este libro. Invito entonces a los lectores a no buscar en él teoremas ni sus demostraciones. Todos los resultados que aparecen en este libro ya fueron demostrados por otros autores. La esencia del texto está en intentar comunicar de la manera más amigable posible, sin perder rigurosidad, un conjunto de técnicas que amplían las ideas generales del análisis de regresión. En cuanto a la rigurosidad, los lectores deben estar preparados en ideas básicas de cálculo, álgebra lineal, inferencia estadística paramétrica y no paramétrica y modelos lineales.

Me he preguntado muchas veces cuáles fueron las razones que me impulsaron a escribir un libro y en particular un libro sobre regresión no paramétrica. En realidad hay tantos libros sobre el tema, que me parece razonable, como se pregunta Simonoff (1996), preguntarme por qué escribir otro más y sobre cuáles serían sus aportes, más allá de lo escrito y publicado.

Una primera respuesta es que este libro está escrito en español. Frederic Ferraty, Vicente Nuñez Antón y Philippe Vieu escribieron su libro *Regresión no Paramétrica: Desde la Dimensión Uno Hasta la Dimensión Infinita* (Ferraty, Nuñez Antón & Vieu 2001) y este es el único que conozco dedicado al tema y escrito en español. Otras notas

de clase en español, no publicadas, se deben al profesor Pedro Delicado de la Universidad Politécnica de Cataluña (Delicado 2008). El libro de Ferraty et. al. está dirigido al capítulo final, ya que su interés principal está en introducir las ideas de regresión con datos funcionales. Tal vez por esta razón su presentación de los estimadores de regresión no paramétrica se limitan a los estimadores kernel. Delicado, por su parte, intenta abarcar desde las pruebas no paramétricas usuales en los cursos de inferencia estadística, pasando por la estimación de densidades para llegar a la estimación no paramétrica, lo que hace su curso extenso deteniéndose menos en los problemas de regresión. Ambos son excelentes textos, pero se parecen poco a la idea plasmada en este, que se concentra en el problema de la regresión no paramétrica vista desde diferentes ángulos.

El libro está organizado en seis capítulos que traté de escribir de tal manera que puedan leerse de manera independiente. El capítulo 1 introduce las ideas generales sobre regresión no paramétrica, tratando de destacar sus similitudes y sus diferencias con el modelo lineal general; el capítulo 2 se adentra en el problema de la regresión no paramétrica, remontándose a sus bases en las series de Fourier; el capítulo 3 presenta los estimadores kernel y el 4 los estimadores spline; el capítulo 5 está dedicado a una breve introducción a los modelos aditivos generalizados; y el capítulo 6 está dedicado al problema de los modelos de regresión en los casos de respuestas múltiples por punto de diseño.

Estoy convencido que nada en la vida se consigue solo. Debo entonces agradecer al profesor K. B. Kulasekera, de la Universidad de Clemson, quien me introdujo a los métodos de regresión no paramétrica hacia finales del siglo XX, y a los profesores del Área de Estadística de la Escuela de Ingeniería Industrial y Estadística de la Universidad del Valle, quienes apoyaron este proyecto de manera irrestricta. Especial gratitud para mis estudiantes de Estadística de la Universidad del Valle, muchos de los cuales construyeron sus trabajos de grado usando estas ideas y fortalecieron mi idea de escribir este libro. Mencionaré especialmente a Alexandra Morales, Martha Montoya, Marco Triana, Cesar Ojeda, Daniel Ayala, Alvaro Florez, Luz Adriana Pereira, Carolina Paz, Andrés Felipe Barrientos, Maria Joana Herrera, Jhovana Reina, Leidy Torres y Diego Castro, cuyos resultados fueron decisivos para entender los elementos que deberían ser incluidos en el libro.

Agradecimientos además a la autoridad municipal (DAGMA) de la ciudad de Santiago de Cali, Colombia, entidad que recolectó y suministró los datos de la mayoría de los ejemplos, datos utilizados originalmente en proyectos del grupo de investigación en Estadística aplicada -INFERIR.

Finalmente, tengo también una deuda de gratitud con la Universidad del Valle, en la que he permanecido mucho más de la mitad de mi vida. La Universidad del Valle me concedió una comisión de estudios, en la cual me familiaricé con los temas de este libro, y luego me concedió un año sabático, en el cual lo escribí. Pero más que eso, la Universidad del Valle ha sido para mí una fuente de vida espiritual y material. Las palabras necesarias para expresar mi gratitud no se han escrito.

El libro se escribió usando el programa de edición de texto LaTeX; las figuras y los programas de cómputo se construyeron usando el software **R**; las figuras se editaron usando Ghostscript y GSView; y las tablas se editaron originalmente usando OpenOffice. Esto significa que todo el software utilizado en la elaboración del libro es de libre distribución. Sin embargo, es menester precisar que todos los programas comerciales modernos de procesamiento estadístico incluyen soluciones para la aplicación del tipo de modelos presentados, incluyendo SAS, SPSS, Stata, Minitab y GenStat, entre otros.

Todo comentario será bienvenido y puede dirigirse a mi dirección de correo-e javier.olaya@correounivalle.edu.co.

Javier Olaya Ochoa

PÁGINA EN BLANCO
EN LA EDICIÓN IMPRESA

Capítulo 1

INTRODUCCIÓN

Según Eubank (1999), cuando alguien dice que ha realizado un análisis de regresión nos imaginamos que esta persona ha estimado unos coeficientes de regresión, ha estimado una varianza, ha construido unos gráficos de residuales, entre otras cosas, como resultado del ajuste de un modelo lineal a un conjunto de datos. En efecto este es el aspecto más conocido del análisis de regresión, aunque hay muchos otros tópicos que merecen incluirse bajo el mismo encabezado. En particular este texto tiene como propósito presentar uno de estos tópicos, relacionado con las ideas asociadas con lo que se conoce como *regresión no paramétrica*.

La expresión *regresión no paramétrica* deja de una vez la sensación, por demás natural, que debe existir una contraparte que se podría llamar *regresión paramétrica*. Parece prudente, entonces, precisar ambas ideas. Previamente, sin embargo, no estaría de más tratar de precisar a qué se refieren otras expresiones tales como *modelo de regresión* o *análisis de regresión*.

El modelo de regresión puede describirse en la siguiente forma general. Supongamos que se desea producir información sobre Y a partir del conocimiento de $\mathbf{x}$. Se dispone de observaciones de la variable continua Y para n valores predeterminados de p variables continuas $X_1, X_2, \ldots, X_p$, las cuales se representan como $\mathbf{x}$. Sean $(\mathbf{x}_i, y_i) = (x_{1i}, x_{2i}, \ldots, x_{pi}, y_i)$, $i = 1, \ldots, n$ las n observaciones de $(\mathbf{x}, \mathbf{Y})$. Se asume que estas observaciones se relacionan a través del

modelo

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

donde $E(\epsilon_i) = 0$ y $var(\epsilon_i) = \sigma^2 < \infty$. Los $f(\mathbf{x}_i)$ son valores de una función desconocida f, conocida como la *función de regresión*, en los puntos $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Y se llama comúnmente la *respuesta*, y $\mathbf{x}$ la *covariable* (o *variable de predicción*, o *variable independiente*, o *variable explicativa*).

El modelo 1.1 provee un marco adecuado para entender dos aproximaciones generales del análisis de regresión: la aproximación *paramétrica* y la aproximación *no-paramétrica*. Estas dos aproximaciones se describen en lo que sigue de este capítulo.

1.1. El modelo de regresión paramétrica: modelos lineales

El caso paramétrico más sencillo, la *regresión lineal simple*, es probablemente el modelo de regresión de uso más común. Según Graybill (1976) un *modelo (estadístico) lineal general* puede escribirse así:

$$y = f(x) + \epsilon, \tag{1.2}$$

donde y y ϵ son variables aleatorias y f es una función de una variable no-aleatoria definida en un dominio D. La *función de regresión* f se considera la porción determinística del modelo. La variable aleatoria ϵ no es observable, pero usualmente se asumen algunas características de su función de densidad.

En los modelos paramétricos se conoce la forma de la función de regresión f, aunque f contiene parámetros desconocidos. En un modelo lineal, f es una función lineal de los parámetros desconocidos. Una de las formas más simples de la función de regresión lineal f es

$$f(x) = \beta_0 + \beta_1 x \tag{1.3}$$

para $x \in D_x$; β_0 y β_1 se definen en un espacio paramétrico Ω_β.

Si utilizamos la función 1.3, podemos escribir de la siguiente manera el modelo 1.2:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1.4}$$

donde $E(\epsilon) = 0$ y $var(\epsilon) = \sigma^2$. En general σ^2 se desconoce. El modelo 1.4 se conoce usualmente como el *modelo lineal simple*.

Como la función de regresión 1.3 se determina completamente con el conocimiento de β_0 y β_1, uno de los propósitos del análisis del modelo lineal simple es estimar β_0 y β_1. Otro objetivo es estimar σ^2, porque las inferencias a partir del modelo 1.4 requieren la estimación de la varianza del error.

Escrito en la forma 1.4, el modelo se llama habitualmente *modelo poblacional*. Se requiere un *modelo muestral* para estimar los parámetros, el cual puede escribirse como el conjunto de ecuaciones

$$\left.\begin{array}{l} y_i = \beta_0 + \beta_1 x_i + \epsilon_i \\ E(\epsilon_i) = 0 \end{array}\right\} \quad i = 1, \ldots, n, \tag{1.5}$$

donde (x_i, y_i), $i = 1, \ldots, n$ son observaciones de la variable X y de la variable aleatoria Y. Los ϵ_i son variables aleatorias no-observables tales que $cov(\epsilon_i, \epsilon_j) = \sigma_{ij}$.

Lo más importante hasta este punto es que los modelos poblacionales 1.2 y 1.4 definen una relación en el conjunto D, el dominio de f, y que Y es una variable aleatoria diferente para cada x en D_x. A partir de este planteamiento, se obtiene una muestra (x_i, y_i), $i = 1, \ldots, n$ y se utilizan estas observaciones para construir inferencias sobre los parámetros desconocidos β_0, β_1.

El modelo lineal general 1.2 puede extenderse al caso de una covariable q-dimensional y se representa así:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

En este caso es preferible escribir el modelo en la notación matricial

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.6}$$

donde $\mathbf{y}$ es un vector aleatorio observable de tamaño $n \times 1$; $\mathbf{X}$ es una matriz $n \times q$ de números dados (los elementos de $\mathbf{X}$ no son variables aleatorias); $\boldsymbol{\beta}$ es un vector $q \times 1$ de parámetros no-observables definidos en un espacio muestral Ω, y $\boldsymbol{\epsilon}$ es un vector aleatorio $n \times 1$ tal que $E(\boldsymbol{\epsilon}) = \mathbf{0}$ y $cov(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}$, siendo $\boldsymbol{\Sigma}$ una matriz definida positiva. q es igual a $p + 1$ en general y en tal caso la primera columna de $\mathbf{X}$ tiene todos sus elementos iguales a 1 y el vector $\boldsymbol{\beta}$ tiene elementos $\beta_0, \beta_1, \ldots, \beta_p$. Las restantes p columnas de $\mathbf{X}$ contienen las observaciones del vector $\mathbf{x}$ de variables continuas independientes.

Supóngase, por ejemplo, que $\mathbf{x} = (X_1, X_2)$. Entonces $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ y $\boldsymbol{\epsilon}$ se definen tal como se presentan en las expresiones 1.7.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{1.7}$$

Si se desea estimar la función de regresión f en el modelo lineal, un método de estimación disponible es el de mínimos cuadrados (Ver Graybill 1976, Neter, Wasserman & Kutner 1990, Seber 1977, Draper & Smith 1998, Searle 1971). Los estimadores de mínimos cuadrados de los parámetros β_j de la expresión 1.7 son los que se ilustran en 1.8.

$$\left.\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}. \end{aligned}\right\} \tag{1.8}$$

donde $\vec{x}$ es la media aritmética de X, definida como

$$\vec{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Así, un estimador $\hat{f}(x)$ de la función de regresión $f(x)$ en el modelo lineal simple 1.4 es

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Si el problema considera un predictor q-dimensional y $(\mathbf{X^T X})^{-1}$ existe, entonces el estimador de mínimos cuadrados $\hat{\boldsymbol{\beta}}$ del vector de parámetros $\boldsymbol{\beta}$ está dado por:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T y} \tag{1.9}$$

donde $\boldsymbol{X}$ y $\boldsymbol{y}$ los hemos definido para el modelo lineal general (1.6). Bajo el supuesto $E(\boldsymbol{\epsilon}) = 0$, $\hat{\boldsymbol{\beta}}$ es un estimador insesgado de $\boldsymbol{\beta}$ y de allí que $\hat{f}$ es un estimador insesgado de f.

Por otra parte, $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X^T X})^{-1}$. Entonces, $\text{var}(\hat{f}) = \text{var}(\boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T}\ \sigma^2 = \boldsymbol{H}\ \sigma^2$, lo cual implica una gran dependencia de $\text{var}(\hat{f})$ en los valores observados de $\mathbf{x}$.

La fórmula (1.9) utilizada para estimar $\boldsymbol{\beta}$ trae consigo una dificultad potencial en el proceso de estimación, ya que si $\boldsymbol{X^T X}$ es singular no disponemos de una solución única. Esto significaría que al menos una columna de $\boldsymbol{X}$ es linealmente dependiente de una o más de las demás columnas. Decimos que existe colinealidad o multicolinealidad entre las columnas de $\boldsymbol{X}$ no solamente si $\boldsymbol{X^T X}$ es singular si no si es en general mal condicionada. El mal-condicionamiento es indeseable en el análisis de regresión paramétrica, porque conduce a estimaciones poco confiables de los parámetros, los que tendrán en ese caso varianzas y covarianzas grandes (Draper & Smith 1998).

En ocasiones se acostumbra ajustar un modelo usando un polinomio de grado p de la forma:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon, \quad i = 1, 2, \ldots, n,$$

que se puede escribir como

$$y_i = \sum_{j=0}^{p} \beta_j x_i^j + \epsilon_i \tag{1.10}$$

Para resolver este problema usando mínimos cuadrados ordinarios, la matriz $\boldsymbol{X}$ tomaría la forma:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^p \\ 1 & x_2 & x_2^2 & \ldots & x_2^p \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^p \end{bmatrix}$$

Otra generalización del modelo (1.2) es considerar el caso de una covariable aleatoria X. En este caso asumimos que disponemos de dos variable aleatorias Z y X cuya densidad conjunta es $\psi_{Z,X}(z, x)$. Sea Y la variable aleatoria $[Z|X = x]$, $E(\epsilon) = 0$ y $\text{var}(\epsilon) = \sigma^2 < \infty$. Diremos que $f(x)$ es el valor esperado de la variable aleatoria $[Z|X = x]$ y por tanto el modelo (1.2) puede escribirse como

$$Y = E(Z|X = x) + \epsilon \tag{1.11}$$

Cuando se asume el modelo (1.11), no se requiere más del modelo muestral (1.5), porque en este caso estamos muestreando una densidad conjunta (Graybill 1976).

Finalmente, podemos tener también un vector aleatorio p-dimensional $\boldsymbol{x}$ como covariable. En este caso asumiremos que disponemos de $p+1$ variables aleatorias $X_0, X_1, \ldots, X_p$, definiremos $Y = [X_0 \mid X_1, \ldots, X_p]$ y asumiremos que la función de regresión $f(\mathbf{x})$ está dada por el valor esperado de la variable aleatoria $[X_0 \mid X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p]$. Bajo este esquema, asumiremos que colectamos n observaciones $\{x_{0i}, x_{1i}, x_{2i}, \ldots, x_{pi}\}$, $i = 1, \ldots, n$ de la densidad conjunta $\psi_{X_0, X_1, \ldots, X_p}(x_0, x_1, \ldots, x_p)$ y asumiremos que el modelo

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

se satisface para las variables aleatorias $Y, X_1, X_2, \ldots, X_p$ y ϵ. Se asume también que $E(\epsilon) = 0$ y $var(\epsilon) = \sigma^2 < \infty$. Finalmente, la distribución de $Y = [X_0 \mid X_1, \ldots, X_p]$ satisface que

$$E[X_0 \mid X_1 = x_1, \ldots, X_p = x_p] = f(x_1, x_2, \ldots, x_p)$$

por lo que en efecto la función de regresión será el valor esperado de la variable aleatoria condicional $[X_0 \mid X_1 = x_1, \ldots, X_p = x_p]$.

Ejemplo 1.1. La Tabla 1.1 consiste en un conjunto de datos de 50 observaciones (x_i, y_i), a las cuales se desea ajustar un modelo lineal.

Tabla 1.1: Datos para el ejemplo 1.1

Obs. #	X	Y	Obs. #	X	Y	Obs. #	X	Y
1	0.01	0.12	18	0.35	0.57	35	0.69	0.82
2	0.03	0.03	19	0.37	0.55	36	0.71	0.52
3	0.05	0.19	20	0.39	0.68	37	0.73	0.67
4	0.07	-0.22	21	0.41	0.40	38	0.75	0.61
5	0.09	-0.24	22	0.43	0.61	39	0.77	1.13
6	0.11	0.28	23	0.45	0.80	40	0.79	1.13
7	0.13	0.01	24	0.47	0.85	41	0.81	0.73
8	0.15	0.10	25	0.49	1.07	42	0.83	0.84
9	0.17	-0.13	26	0.51	0.78	43	0.85	0.76
10	0.19	0.49	27	0.53	1.29	44	0.87	0.97
11	0.21	-0.02	28	0.55	0.82	45	0.89	0.69
12	0.23	0.42	29	0.57	0.82	46	0.91	0.59
13	0.25	0.32	30	0.59	1.20	47	0.93	1.10
14	0.27	0.20	31	0.61	0.39	48	0.95	1.26
15	0.29	0.56	32	0.63	0.74	49	0.97	0.74
16	0.31	0.37	33	0.65	0.81	50	0.99	0.93
17	0.33	0.66	34	0.67	1.00			

Solución: Como el investigador conoce el origen de sus datos, quizás tenga en mente algún modelo lineal que puede anticipar como adecuado para estudiar conjuntamente estas dos variables. En ausencia de esta información, los analistas acuden a algunas técnicas preliminares para intentar hacerse a una idea sobre la función de

regresión subyacente a estos datos. Una estrategia común es construir un diagrama de puntos como el que se muestra en la figura 1.1.

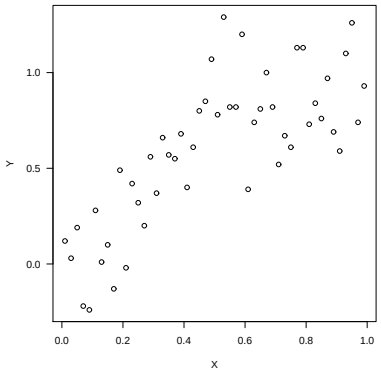


Figura 1.1: Diagrama de puntos de los datos de la Tabla 1.1

El diagrama de puntos de la figura 1.1 no es muy informativo sobre cuál modelo podría ajustarse a los datos. Una posible salida en este caso es realizar una búsqueda de posibles modelos, evaluando su bondad para decidir si alguno de ellos es razonablemente aceptable para efectos del estudio en curso. En este ejemplo se podría utilizar esta estrategia tal como se ilustra en la figura 1.2, que muestra cuatro veces el diagrama de puntos del conjunto de datos de la tabla 1.1. En cada uno de los diagramas se ha dibujado un modelo lineal diferente para estimar la función de regresión. El modelo lineal simple de la parte superior izquierda de la figura 1.2 no supera los diagnósticos del modelo, que indican que posiblemente se requiera un término no-lineal en X. El modelo cuadrático de la parte superior derecha tampoco supera los diagnósticos de los supuestos del modelo, indicando que posiblemente se necesite un término no-lineal en X mayor que el cuadrático. En el lado inferior izquierdo de la figura 1.2 se dibuja un polinomio de grado 4 ajustado a los datos, cuyo término lineal no es significativo, por lo que se suprime del modelo, para ajustar un polinomio de grado 4 sin el término lineal en X. Este último polinomio se dibuja en la parte inferior derecha de la figura 1.2.

Los cuatro modelos ajustados aparecen en las ecuaciones 1.12

$$\left.\begin{aligned} \hat{f}(x) &= 0.0926 + 1.0152x \\ \hat{f}_2(x) &= -0.1719 + 2.6019x - 1.5867x^2 \\ \hat{f}_{41}(x) &= 0.0813 - 2.3637x + 20.4974x^2 - 34.0781x^3 + 16.9396x^4 \\ \hat{f}_{42}(x) &= -0.0666 + 11.1891x^2 - 20.8344x^3 + 10.7287x^4 \end{aligned}\right\} \tag{1.12}$$

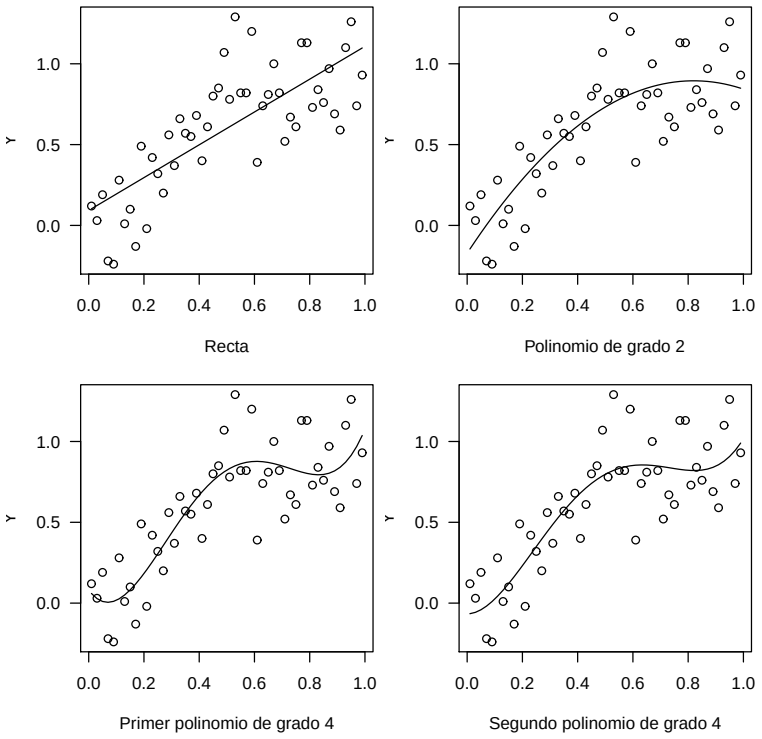


Figura 1.2: Cuatro modelos lineales ajustados a los datos de la tabla 1.1: modelo lineal simple (arriba-izquierda); modelo parabólico (arriba-derecha); polinomio de grado 4 (abajo-izquierda); y polinomio de grado 4 sin el término lineal (abajo-derecha)

En todo caso, la meta será estimar un número finito de parámetros de una función cuya forma se ha establecido previamente. □

Como $\hat{f}$ es un estimador de f, una pregunta propia del análisis de regresión es cuáles son sus propiedades, para efectos de inferencia.

Por lo pronto, se sabe que $\hat{f}(x)$ es un estimador *insesgado* de $f(x)$, es decir, $E[\hat{f}(x)] = f(x)$, asumiendo que la verdadera función de regresión f es lineal. Este resultado sigue del hecho que $\hat{\beta}_0$ es un estimador insesgado de β_0 y $\hat{\beta}_1$ es un estimador insesgado de β_1.

Además, la varianza de $\hat{f}(x)$ está dada por la ecuación 1.13, donde σ^2 es la varianza de los errores en el modelo 1.4.

$$Var(\hat{f}(x)) = \sigma^2 \left[n^{-1} + (x - \bar{x})^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2 \right] \tag{1.13}$$

Por otra parte, $\hat{f}$ es un estimador *consistente* de f, cuyo error cuadrático medio converge a cero a una tasa de n^{-1} (ver ejercicio 1).

Se puede afirmar en consecuencia, al menos basados en estas dos propiedades y bajo el supuesto citado de que la función de regresión f es lineal, que $\hat{f}$ es un *buen* estimador de f.

1.2. El modelo de regresión no paramétrica

Los objetivos del análisis de regresión no paramétrica son los mismos de su contraparte paramétrica, vale decir, estimar y probar las características de la función de regresión.

En el análisis de regresión se parte de la idea de que existe una función de regresión f subyacente, desconocida. Tal como en los modelos lineales, en el modelo de regresión no paramétrica será necesario establecer algunas suposiciones acerca de f. Por ejemplo, uno podría exigir como parte del modelo (1.1) que la función de regresión sea continua (tal como lo es en el modelo lineal). Pero además podría ser necesario garantizar que f tenga un número determinado, digamos λ, de derivadas continuas.

1.2.1. Suavización

Supongamos que disponemos de una variable explicativa simple X y de una respuesta Y y que el modelo (1.1) es aplicable al par (X, Y). Sea f una función perteneciente a algún espacio de funciones W. La selección de W depende de qué tan suave es posible asumir f. Usualmente el espacio W se asume como el conjunto de todas las funciones f que son continuas en un intervalo cerrado $[a, b]$, o quizás como el conjunto de todas las funciones f que tienen k derivadas continuas en (a, b), para algún entero positivo k. Los límites los fija el investigador, pero se puede asumir, sin pérdida de generalidad, que son $a = 0$ y $b = 1$.

El procedimiento para estimar la función de regresión f en el marco del análisis de regresión no-paramétrica se llama *suavización*. La siguiente es la idea básica de *suavización*, siguiendo a Eubank (1999, pg. 12): Se asume que la función de regresión f es suave y que se desea estimar f en el punto x. Entonces, debido a la suavidad de f, las observaciones y_i obtenidas en los puntos x_i cercanos a x deben contener información acerca del valor de f en el punto x. Por lo tanto, debe ser posible utilizar algún tipo de promedio de las respuestas y_i en aquellos puntos x_i cercanos a x para estimar $f(x)$.

Sea f una función suave y supóngase que se desea estimar f en el punto x. Se dispone de observaciones de f afectadas por cierto error aleatorio (nuestras y_i) en algunos puntos x_i cercanos a x. Entonces las observaciones y_i en esos puntos x_i cercanos a x deben contener información cerca de f en el punto x debido a la suavidad de f. Así, debe ser posible utilizar algún tipo de promedio de las respuestas y_i en esos puntos x_i cercanos a x para obtener una estimación de $f(x)$.

Un *suavizador* es una herramienta para resumir el comportamiento de una respuesta Y como función de una variable explicativa X. X puede ser un vector aleatorio. Los suavizadores más comunes se conocen como *estimadores lineales* de la función de regresión y tiene la forma general dada en la ecuación 1.14.

$$\hat{f}(x) = \sum_{i=1}^{n} K(x, x_i; \lambda) y_i \tag{1.14}$$

donde $K(x, x_i; \lambda)$, $i = 1, \ldots, n$, es una colección de pesos que dependen del diseño $\{x_i,\ i = 1, \ldots, n\}$ y de un parámetro λ definido por el usuario. Nótese que $\hat{f}$ depende de λ, aunque esta dependencia no se hace explícita en la notación utilizada hasta aquí. Una notación que remarcaría esta dependencia sería escribir f_λ en lugar de $\hat{f}$. En todo caso, debe ser claro en el contexto que $\hat{f}$ depende de λ. Estos estimadores se llaman "lineales" porque para un λ dado, los estimadores resultan ser funciones lineales de las respuestas y_i.

Si los pesos $K(x, x_i; \lambda)$, $i = 1, \ldots, n$ son tales que todos son no-negativos y que $\sum_{i=1}^{n} K(x, x_i; \lambda) = 1$, entonces el suavizador (1.14) es un promedio ponderado de las respuestas Y_i. Un ejemplo de un suavizador de este tipo es el estimador de Nadaraya-Watson (Nadaraya & Seckler 1964) & (Watson 1964):

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K(\lambda^{-1}(x - x_i)) y_i}{\sum_{i=1}^{n} K(\lambda^{-1}(x - x_i))}. \tag{1.15}$$

El estimador de Nadaraya-Watson pertenece a una clase de estimadores lineales conocidos como estimadores *kernel*. Los pesos en un estimador kernel se derivan de una función kernel común que no depende del conjunto de valores de X que intervienen en la regresión (a este conjunto de valores de X se le conoce como *el diseño*). Se dice entonces que los pesos en un estimador kernel no dependen del diseño. La función kernel se denota usualmente K y en general es una

función de densidad de probabilidad simétrica alrededor de cero y cuyo pico ocurre en cero. La función de asignación de pesos decae a cero a medida que se aleja del punto de estimación x.

En un estimador kernel los pesos toman la forma dada en la ecuación 1.16.

$$K(x, x_i; \lambda) = \frac{1}{\lambda} K(\frac{x - x_i}{\lambda}) \tag{1.16}$$

Si la función K es simétrica y apuntada en cero, entonces y_i contribuye más a la estimación de $f(x)$ en el punto x_i asociado con tal y_i.

Dos ejemplos de funciones kernel son el kernel *uniforme*

$$K(u) = \frac{1}{2} I_{[-1,1]}(u)$$

y el kernel *cuadrático*

$$K(u) = \frac{3}{4}(1 - u^2) I_{[-1,1]}(u).$$

El parámetro λ en (1.14) se llama el *ancho de banda* ó *parámetro de suavización*, el cual puede ser cualquier número no-negativo.

Un estimador lineal que utilizara pesos como los definidos en (1.16) sería:

$$\hat{f}(x) = (n\lambda)^{-1} \sum_{i=1}^{n} K(\lambda^{-1}(x - x_i)) y_i.$$

El estimador de Nadaraya-Watson (1.15) es una variación de esta forma general en el cual los pesos han sido normalizados.

Otra forma posible es el estimador de Gasser-Müller (Gasser & Müller 1979) & (Müller 1988)

$$\hat{f}(x) = \sum_{i=1}^{n} [\lambda^{-1} \int_{s_{i-1}}^{s_i} K(\lambda^{-1}(x - s)) ds] y_i, \tag{1.17}$$

donde $s_0 = 0$, $s_{i-1} \leq x_i \leq s_i$, $i = 1, \ldots, n$, $s_n = 1$.

Los estimadores de Nadaraya-Watson y de Gasser-Müller son estimadores consistentes pero no insesgados de $f(x)$. Por ejemplo el sesgo y la varianza del estimador de Gasser-Müller (1.17) son, respectivamente,

$$E[\hat{f}(x)] - f(x) = \frac{\lambda^2}{2} f''(x) M_2 + o(\lambda^2) + O(n^{-1}),$$

donde $M_2 = \int_{-1}^{1} u^2 K(u)du \neq 0$, y

$$var[\hat{f}(x)] = \frac{\sigma^2 V}{n\lambda} + O((n\lambda)^{-2}), \tag{1.18}$$

con $V = \int_{-1}^{1} K(u)^2 du$. K se escoge tal que $V < \infty$. Por lo tanto se requeriría $n \to \infty$, $\lambda \to 0$ y $n\lambda \to \infty$ para que el sesgo y la varianza del estimador de Gasser-Müller (1.17) converjan a cero.

Otro estimador común, que también es un estimador kernel, es el estimador de regresión local (Cleveland 1979), el cual se ha popularizado con el nombre de estimador LOESS, el más simple de los cuales utiliza los k vecinos más cercanos y estima el valor de la función de regresión en el punto x conforme a los siguientes pasos:

1 Identificar los k vecinos más cercanos de x y denotar este conjunto como $N(x)$

2 Encontrar $\Delta(x) = max_{\{x_i \in N(x)\}}|x - x_i|$ (la distancia asociada con el vecino mas cercano que se encuentra más alejado de x).

3 Asignar pesos w_i a cada punto en $N(x)$ usando la función de pesos tri-cubo

$$\kappa\left(\frac{|x - x_i|}{\Delta(x)}\right)$$

donde

$$\kappa(u) = \begin{cases} (1-u^3)^3, & \text{para } 0 \leq u \leq 1 \\ 0, & \text{en cualquier otro caso} \end{cases}$$

4 Ajustar una recta por mínimos cuadrados de Y sobre X, confinada al conjunto $N(x)$, utilizando los pesos obtenidos en 3.

La estimación lineal local LOESS de $f(x)$ en el punto x toma como valor el del término independiente de la la recta de regresión local resultante. Es decir, si la recta ajustada es $\hat{\beta}_o + \hat{\beta}_1 x$, entonces $\hat{f}(x) = \hat{\beta}_o$.

Estos y otros estimadores comunes en el modelo de regresión no-paramétrica se discutirán con más detalle en capítulos posteriores.

Ejemplo 1.2. Ajustar un modelo de regresión no paramétrico a los datos de la tabla 1.1.

Solución: En este caso no se requiere anticipar la forma de la función de regresión, por lo que se dice que una característica de los modelos de regresión no paramétrica es "dejar que los datos *hablen*". La figura 1.3 muestra dos posibles ajustes no paramétricos: un ajuste de Nadaraya-Watson con $\lambda = 0.45$ y un ajuste LOESS con parámetro de suavización (span) de 0.75.

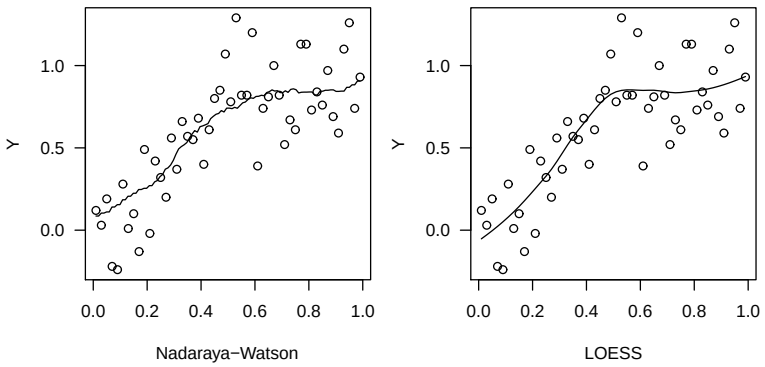


Figura 1.3: Modelo de Nadaraya-Watson (izquierda) y modelo LOESS (derecha) ajustados al conjunto de datos de la tabla 1.1

Nótese que el ajuste LOESS es "más suave" que el de Nadaraya-Watson, con menos fluctuaciones.

□

Una pregunta que surge naturalmente de las soluciones a los ejemplos es: ¿Cuál de todos estos modelos (lineales del ejemplo 1.1 y no paramétricos del ejemplo 1.2) estima mejor la función de regresión? En las aplicaciones, esta pregunta tropieza con la dificultad de que uno no conoce la función de regresión. Así que se acude a las propiedades de los estimadores para decidir cuál modelo elegir, en el proceso que habitualmente se conoce como *análisis de regresión*. Las propiedades de los estimadores en los modelos lineales es un problema bien estudiado en la literatura (Graybill 1976) & (Neter et al. 1990) & (Seber 1977) & (Draper & Smith 1998) & (Searle 1971) y las propiedades de los estimadores en regresión no paramétrica se presentarán en el capítulo 2.

Para intentar una primera comparación, en los ejemplos 1.1 y 1.2, los datos de la tabla 1.1 se generan a partir de la función $f(x) = x + 0.5e^{-50(x-0.5)^2}$, de tal manera que los valores de X son puntos igualmente espaciados $x_i = (2i-1)/100,\ i = 1, \ldots, 50$ y

los valores de Y resultan de añadir 50 números aleatorios generados de una distribución normal con media 0 y varianza 0.2 a $f(x_i) = x_i + 0.5 * exp\{(-50 * (x_i - 0.5)^2)\},\ i = 1, \ldots, 50$, conservando los dos primeros decimales de los valores de Y. Es decir, en este problema se conoce la función de regresión.

Ejemplo 1.3. Comparar gráficamente los modelos lineales ajustados en el ejemplo 1.1 con los modelos no paramétricos ajustados en el ejemplo 1.2 utilizando la función de regresión $f(x) = x + 0.5e^{-50(x-0.5)^2}$.

Solución: La figura 1.4 muestra las comparaciones con la función de regresión. El modelo cuadrático no se desempeña muy bien, mientras que los restantes tres modelos recuperan bastante bien el comportamiento de la función de regresión. En particular, el modelo LOESS parece reconstruir mejor que los demás la función de regresión especialmente para los valores de X por debajo de 0.5. El modelo polinómico de grado cuatro sin la componente lineal también se desempeña muy bien. La diferencia entre el modelo lineal polinómico de grado cuatro y el modelo no paramétrico LOESS es que en el primer caso el investigador debe llegar o bien a predeterminar que este es el mejor modelo para sus datos o bien a buscar mediante una exploración sistemática que éste sería un buen modelo, mientras que en el segundo caso el método "extrae" de los datos un comportamiento posible.

Se destaca también que el ajuste de Nadaraya-Watson, aunque sigue la forma general de la función de regresión, no tiene la suavidad de los otros modelos. Esta es una característica del estimador de Nadaraya-Watson, el primer modelo formulado explícitamente entre los modelos conocidos de regresión no paramétrica.

La Figura (1.5) muestra intervalos del 95 % de confianza punto-a-punto para los modelos polinómico de grado 4 (izquierda) y LOESS (derecha).

□

1.2.2. Algunos Resultados de Inferencia en el Modelo No-Paramétrico

De acuerdo con la discusión anterior, se dispone de una variedad de opciones de entre las cuales es posible elegir un estimador $\hat{f}$ para la función de regresión f en el marco no-paramétrico. Naturalmente sería preferible escoger un "buen" estimador. Para lograr un buen estimador se requiere algún criterio de "bondad" de tales estimadores junto a una

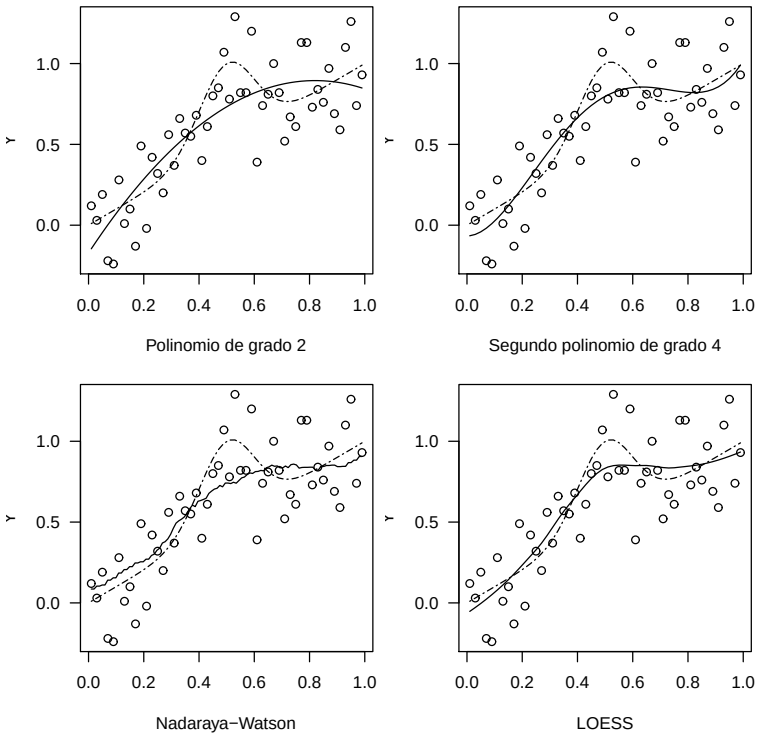


Figura 1.4: Comparación con la función de regresión de los modelos cuadrático (arriba-izquierda) y polinomial de grado 4 (arriba-derecha) del ejemplo 1.1 y de los modelos de Nadaraya-Watson (abajo-izquierda) y LOESS (abajo-derecha)

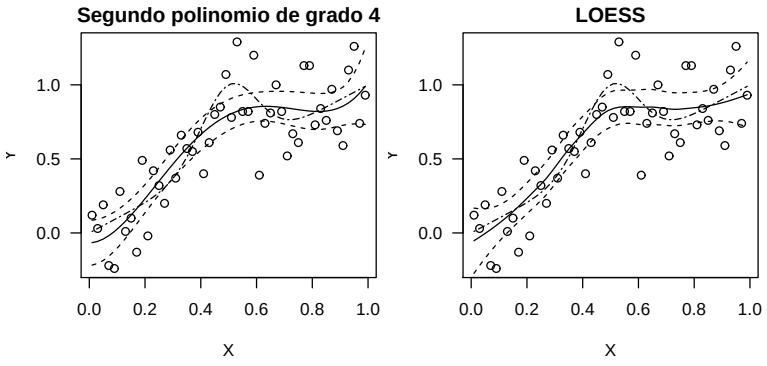


Figura 1.5: Intervalos del 95 % de confianza punto-a-punto para los modelos polinómico de grado 4 (izquierda) y LOESS (derecha)

herramienta para medirla. Un indicador aceptado ampliamente es el *riesgo*, definido como:

$$R(\hat{f}) = E[f(x) - \hat{f}(x)]^2.$$

Lo deseable en este caso es obtener un estimador $\hat{f}$ de f que tenga un riesgo tan bajo como sea posible. El riesgo se puede separar en dos componentes:

$$R(\hat{f}) = B^2(\hat{f}) + var(\hat{f}),$$

donde $B(\hat{f})$ es el sesgo del estimador $\hat{f}$ y $var(\hat{f})$ es su varianza.

Observando el sesgo y la varianza del estimador de Gasser-Müller (ver ecuaciones 1.18), se destaca una de las propiedades del parámetro de suavización λ en la estimación no-paramétrica. Para un n dado y $\lambda < 1$, si λ es pequeño el sesgo será pequeño, pero la varianza será grande. En general, anchos de banda pequeños hacen que los estimadores lineales tengan varianzas altas junto con sesgos pequeños, y viceversa. Este hecho impacta en alto grado las estimaciones debido a que anchos de banda pequeños producirán estimaciones muy variables, mientras que anchos de banda grandes producirán estimaciones más suaves pero muy sesgadas.

Supóngase que se dispone de dos estimadores $\hat{f}_1$ y $\hat{f}_2$ de una función de regresión f. Más aún, asúmase que ambos estimadores tienen asociado el mismo riesgo, es decir,

$$R(\hat{f}_1) = B^2(\hat{f}_1) + var(\hat{f}_1) = B^2(\hat{f}_2) + var(\hat{f}_2) = R(\hat{f}_2)$$

Entonces, si $\hat{f}_1$ es menos sesgado que $\hat{f}_2$, se sigue que $\hat{f}_1$ tiene mayor varianza que $\hat{f}_2$.

Hasta este momento se ha asumido que la estimación no-paramétrica se basa en un diseño no aleatorio. Un diseño no aleatorio es un conjunto de valores de X, por ejemplo $\{x_i,\ i = 1, \ldots, n\}$, el cual se fija por anticipado. Se observan las respuestas Y_i en estos puntos de diseño. Sin embargo, es posible definir estimadores no-paramétricos para el caso de variables explicativas aleatorias. De hecho, el estimador de Nadaraya-Watson fue introducido inicialmente para el caso de variables explicativas aleatorias (Nadaraya & Seckler 1964) y extendido luego al caso de diseños no aleatorios (Benedetti 1975).

Puede demostrarse (Eubank 1999, pg. 213) que el estimador de Nadaraya-Watson y el estimador de Gasser-Müller tienen los mismos

términos principales (ver ecuaciones 1.18) en las expresiones para el sesgo y la varianza si el diseño no es aleatorio. Sin embargo, esto no ocurre cuando los x_i son aleatorios (Eubank 1999, pg. 212). Se sabe (Eubank 1999, pg. 213) que el estimador de Gasser-Müller tiene el mismo sesgo con diseños aleatorios o no aleatorios. Pero su varianza se multiplica por un factor de 1.5 si el diseño es aleatorio. Entretanto, el estimador de Nadaraya-Watson tiene igual varianza en ambos casos y su sesgo depende de la densidad de X en el caso aleatorio. Como los dos estimadores tienen varianzas y sesgos asintóticos que dependen del tipo de diseño, la pregunta sobre cuál tipo de diseño es preferible en el caso de x_i's aleatorios se vuelve importante. La alta varianza del estimador de Gasser-Müller lo hace ciertamente indeseable, pero al mismo tiempo su sesgo es más simple y menor para algunas densidades de X que el sesgo del estimador de Nadaraya-Watson. Se diría que un estimador "ideal" debería tener la varianza del estimador de Nadaraya-watson y el sesgo del estimador de Gasser-Müller. Pues bien, el estimador LOESS (Eubank 1999, pg. 214) posee ambas características. Parece entonces recomendable el uso de estimadores LOESS cuando se tiene el caso de diseños aleatorios.

1.2.3. La maldición de la dimensionalidad

Es posible generalizar la estimación no-paramétrica al caso p-dimensional. En primer lugar es necesario enfatizar tres dificultades que surgen en en el problema de la estimación no-paramétrica multivariante:

1 Los estimadores multivariantes son necesariamente más complicados que los univariados. En la práctica hay muchas más opciones que elegir y se deben escoger más parámetros de suavización.

2 La visualización gráfica es difícil en la estimación multivariante. Para el caso de una variable explicativa bidimensional es aún posible obtener alguna visualización. Pero ciertamente no lo es para variables explicativas en tres o más dimensiones.

3 A medida que la dimensión de la variable explicativa se incrementa, los estimadores son cada vez más inexactos. Supóngase que se alcanza una exactitud dada en la estimación de una función de regresión en un espacio k-dimensional. Entonces,

en dimensiones mayores que k, se necesitan muestras mucho más grandes para alcanzar la misma exactitud. Este hecho se conoce como *la maldición de la dimensionalidad*. Una consecuencia de ello es que, en dimensiones altas, vecindades "locales" están en casi todas vacías y vecindades que no están vacías, no son en general "locales".

Considérese el siguiente ejemplo (Simonoff 1996). Supóngase que se tiene una muestra uniforme en el hipercubo $[-1, 1]$. 79 % de las observaciones caerán en el círculo unitario centrado en el origen cuando $p = 2$, pero solamente 15 % si $p = 5$ y 0.25 % si $p = 10$. Es decir, vecindades grandes prácticamente no incluyen observaciones, implicando la pérdida del caracter local de la estimación.

La forma general de un estimador kernel multidimensional es

$$f_L(\mathbf{x}) = \frac{1}{n|\mathbf{L}|} \sum_{i=1}^{n} K_p[\mathbf{L}^{-1}(\mathbf{x} - \mathbf{x}_i)] \tag{1.19}$$

donde $K_p : \Re^p \to \Re$ es la función kernel (la cual es en general una función de densidad de probabilidad); $\mathbf{L}$ es una matriz $p \times p$ de anchos de banda, no-singular; y $|\mathbf{L}|$ es el valor absoluto del determinante de la matriz $\mathbf{L}$. Una técnica común para obtener K_p a partir de una función kernel univariada K es utilizar un producto de la forma $K_p(\mathbf{u}) = \prod_{j=1}^{p} K(u_j)$.

La generalización del estimador LOESS al caso p-dimensional no es difícil, porque la estimación LOESS de la función de regresión en el punto x_0 coincide con la estimación de β_0 en el modelo p-dimensional ajustado en una vecindad k-NN alrededor de $X = x_0$ (Simonoff 1996). El estimador $\hat{f}(x_0)$ es el elemento $\hat{\beta}_0$ del vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ que minimiza

$$\sum_{i=1}^{n} [y_i - \beta_0 - \beta_1(x_i - x_{1i}) - \ldots - \beta_p(x_p - x_{pi})]^2 K_p[\mathbf{L}^{-1}(\mathbf{x} - \mathbf{x}_i)]$$

donde $\mathbf{L}$ y K_p se definen como en (1.19).

1.3. Un llamado a la mesura

Los métodos de regresión no paramétrica, con sus estimadores no insesgados, no son un substituto de los modelos lineales, con

sus estimadores máximo verosímiles insesgados de varianza mínima, en los casos en los que los supuestos de los modelos lineales se satisfacen plenamente. De hecho, los métodos de suavización propios de la regresión no paramétrica bien podrían ser un paso previo para la selección de un modelo lineal adecuado para un problema en particular. En consecuencia, los métodos de regresión no paramétrica no se presentan en este texto como un reemplazo per sé de los modelos lineales, si no como una opción teóricamente sustentada para los casos en los que no hay defensa para la linealidad.

1.4. ¿Qué sigue?

Luego de esta rápida visión, se presentará en el capítulo 2 un primer estimador de la función de regresión usando series de Fourier, para pasar naturalmente a los estimadores kernel en al capítulo 3 y a los estimadores spline en el capítulo 4. El capítulo 5 se ocupará de los modelos aditivos generalizados y el capítulo 6 se dedicará al problema de la estimación en problemas de medidas repetidas.

Tabla 1.2: Mediciones en niños con Diabetes Mellitus tipo I (dependiente de insulina)

Obs. #	Edad	Deficit base	Péptido C	Obs. #	Edad	Deficit base	Péptido C
1	5.2	-8.1	4.8	23	11.3	-3.6	5.1
2	8.8	-16.1	4.1	24	1	-8.2	3.9
3	10.5	-0.9	5.2	25	14.5	-0.5	5.7
4	10.6	-7.8	5.5	26	11.9	-2	5.1
5	10.4	-29	5	27	8.1	-1.6	5.2
6	1.8	-19.2	3.4	28	13.8	-11.9	3.7
7	12.7	-18.9	3.4	29	15.5	-0.7	4.9
8	15.6	-10.6	4.9	30	9.8	-1.2	4.8
9	5.8	-2.8	5.6	31	11	-14.3	4.4
10	1.9	-25	3.7	32	12.4	-0.8	5.2
11	2.2	-3.1	3.9	33	11.1	-16.8	5.1
12	4.8	-7.8	4.5	34	5.1	-5.1	4.6
13	7.9	-13.9	4.8	35	4.8	-9.5	3.9
14	5.2	-4.5	4.9	36	4.2	-17	5.1
15	0.9	-11.6	3	37	6.9	-3.3	5.1
16	11.8	-2.1	4.6	38	13.2	-0.7	6
17	7.9	-2	4.8	39	9.9	-3.3	4.9
18	11.5	-9	5.5	40	12.5	-13.6	4.1
19	10.6	-11.2	4.5	41	13.2	-1.9	4.6
20	8.5	-0.2	5.3	42	8.9	-10	4.9
21	11.1	-6.1	4.7	43	10.8	-13.5	5.1
22	12.8	-1	6.6				

1.5. Ejercicios

1. Verifique que en el modelo lineal simple se satisface que:

a. Si la función de regresión f es en realidad lineal, entonces $\hat{f}$ es un estimador insesgado de f, es decir, se cumple que $E[\hat{f}(x)] = f(x)$.

b. La varianza de $\hat{f}$ está dada por la expresión:

$$Var[\hat{f}(x)] = \sigma^2 \left[n^{-1} + (x - \bar{x})^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]$$

c. Si los valores de X son puntos igualmente espaciados $x_i = \frac{2i-1}{2N},\ i = 1, \ldots, N$, entonces $\hat{f}$ es un estimador consistente de f cuyo error cuadrático medio (riesgo) converge a cero a una tasa n^{-1}.

2. Los datos de la tabla 1.2 se refieren a mediciones en sangre de una sustancia llamada péptido C, un subproducto que se crea cuando se produce la hormona insulina. El objetivo del estudio era estudiar la dependencia del nivel sérico de péptido C a partir de otras mediciones, con el fin de ganar en la comprensión del comportamiento de la secreción de insulina residual. Se dispone de observaciones en 43 pacientes de la edad (en meses) y de una medida de acidez llamada deficit base. Las mediciones de Péptido C están dadas como el logaritmo de la concentración en $pmol/ml$ y la edad se da en meses. El estudio es de Sockett, Daneman, Clarson & Ehrich (1987) y los datos han sido tomados de Hastie & Tibshirani (1990, Pg. 304).

a. Ajuste un modelo de regresión lineal simple para estimar el nivel sérico de péptido C en sangre a partir de la edad. Indique por qué es un buen modelo o explique por qué no lo es.

b. Ajuste un modelo de regresión lineal múltiple para estimar el nivel sérico de péptido C en sangre a partir de la edad y del deficit base de los participantes en el estudio. Indique por qué es un buen modelo o explique por qué no lo es.

Capítulo 2

UN ESTIMADOR DE LA FUNCIÓN DE REGRESIÓN

2.1. Introducción

Para establecer un marco general, asumiremos que disponemos de observaciones de la variable de respuesta Y para n valores predeterminados de una variable independiente X. Las n observaciones bivariadas disponibles, denotadas $(x_1, y_1), \ldots, (x_n, y_n)$, siguen el modelo

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{2.1}$$

donde $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ es un vector de errores aleatorios no correlacionados que tienen media cero y varianza común σ^2, f es una función de regresión desconocida y se satisface que $0 \leq x_1 < \cdots < x_n \leq 1$.

Para efectos de presentación de los resultados, asumiremos que los valores de X han sido elegidos así:

$$x_i = (2i - 1)/2n, \quad i = 1, 2, \ldots, n \tag{2.2}$$

Nuestro propósito es estimar f en 2.1, para lo cual buscaremos construir estimadores lineales que puedan escribirse en la siguiente forma general, que para un λ dado es una combinación lineal de las observaciones y_i, donde $K(\cdot, x_i; \lambda)$, $i = 1, \ldots, n$ es una colección de funciones ponderadoras que dependen de los x_i y de un parámetro de

suavización denotado λ:

$$f_\lambda(x) = \sum_{i=1}^{n} K(x, x_i; \lambda) y_i \tag{2.3}$$

2.2. Estimador con series de Fourier

Al final del siglo XVIII, Fourier propuso un método para aproximar funciones periódicas usando combinaciones lineales de funciones trigonométricas sencillas. Sin embargo, aunque el método propuesto por Fourier se refiere a funciones continuas, sus ideas se pueden aplicar mejor en el contexto más amplio de las funciones cuadrado integrables (Arango & Calderón 2006). Si nos restringimos al intervalo $[0, 1]$, este espacio de las funciones cuadrado integrables se denota en general como $L_2[0, 1]$ (Kreyszig 1989, Bartle 1995). El espacio $L_2[0, 1]$, el cual satisface las propiedades generales de un espacio vectorial (Kreyszig 1989, Pág. 53), se define más formalmente en el Anexo 1 (página 42).

Asumamos entonces que $f \in L_2[0, 1])$ y que puede representarse usando la expansión

$$\sum_{j=1}^{\infty} \beta_j f_j \tag{2.4}$$

donde los $\beta_j,\ j = 1, 2, \ldots$ son *coeficientes generalizados de Fourier*, definidos en en el Anexo 1 (página 43) y las funciones $f_j,\ j = 1, 2, \ldots$ conforman una colección de funciones en $L_2[0, 1]$.

Diremos que la colección de funciones $f_j,\ j = 1, 2, \ldots$ utilizada para construir la expansión 2.4 es una *sucesión ortonormal completa CONS*, que definimos formalmente en el Anexo 1 (página 43).

En consecuencia, el modelo 2.1 puede representarse como:

$$y_i = \sum_{j=1}^{\infty} \beta_j f_j(x_i) + \epsilon_i,\ i = 1, 2, \ldots, n \tag{2.5}$$

Lo que significa que los datos siguen un modelo lineal con infinitos coeficientes de regresión desconocidos.

Ahora bien, un resultado conocido sobre los coeficientes de Fourier, llamado Relación de Parseval ($\sum_{j=1}^{\infty} \beta_j^2 = \|f\|^2$, Anexo 1, página 42), nos permite concluir que los β_j decaerán a cero en algún momento.

Uno podría entonces asumir que existe un entero λ tal que

$$f \doteq \sum_{j=1}^{\lambda} \beta_j f_j$$

y por tanto que podríamos escribir la aproximación

$$y_i \doteq \sum_{j=1}^{\lambda} \beta_j f_j(x_i) + \epsilon_i, \ i = 1, 2, \ldots, n \tag{2.6}$$

Pero este modelo 2.6 luce tal como un modelo lineal, por lo que una posible solución al problema de la estimación de f sería estimar los coeficientes de Fourier $\{\beta_j\}_{j=1}^{\lambda}$ usando la técnica de mínimos cuadrados, para lo cual definiremos la matriz $\boldsymbol{X}_\lambda = \{f_j(x_i)\}_{i=1,2,\ldots,n;\ j=1,2,\ldots,\lambda}$. Si $(\boldsymbol{X}_\lambda^T \boldsymbol{X}_\lambda)^{-1}$ existe, entonces el estimador de β tendrá la forma general:

$$\boldsymbol{\beta}_\lambda = (\boldsymbol{X}_\lambda^T \boldsymbol{X}_\lambda)^{-1} \boldsymbol{X}_\lambda^T \boldsymbol{y} \tag{2.7}$$

con $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$.

Entonces nuestro estimador de f será:

$$f_\lambda(x) = \sum_{j=1}^{\lambda} \beta_{\lambda j} f_j(x) \tag{2.8}$$

Supongamos que decidimos utilizar este estimador de f. Quedarían en ese caso algunas dudas por resolver, recordando que $\sum_{j=1}^{\lambda} \beta_j f_j$ es solamente una aproximación a f y que la verdadera función de regresión es en realidad de dimensión infinita:

1. ¿Cómo se estima la varianza σ^2?
2. ¿Como se podría utilizar este modelo ajustado para realizar inferencias?
3. ¿Es el estimador 2.8 un estimador consistente? ¿Qué tan bueno es este estimador?
4. ¿Cómo se elige λ?

En lo que sigue propondremos algunas posibles respuestas a estas preguntas.

2.2.1. Estimación de la varianza

Una vez estimada la función de regresión, el análisis de regresión requiere estimar la varianza σ^2 para poder continuar con el proceso de inferencia estadística. Un primer acercamiento podría ser adaptar el estimador de varianza en el análisis de regresión lineal:

$$\sigma^2_\lambda = \frac{SCE_\lambda}{(n-\lambda)} \tag{2.9}$$

donde la suma de cuadrados de los residuales SCE_λ se define como:

$$SCE_\lambda = \sum_{i=1}^{n}(y_i - f_\lambda(x_i))^2 \tag{2.10}$$

Sin embargo esta definición no está exenta de dificultades, porque el estimador σ^2_λ que hemos construido en 2.9 depende de un valor desconocido λ que ya se ha utilizado para estimar f. A diferencia del estimador de varianza en el modelo de regresión lineal, el estimador σ^2_λ (definido en 2.9), no es en general insesgado. El sesgo del estimador σ^2_λ es:

$$B(\sigma^2_\lambda) = \frac{\boldsymbol{f}^T(\boldsymbol{I} - \boldsymbol{S}_\lambda)\boldsymbol{f}}{n-\lambda} \tag{2.11}$$

con $\boldsymbol{f} = (f(x_1), f(x_2), \ldots, f(x_n))^T$ y $\boldsymbol{S}_\lambda = \boldsymbol{X}_\lambda(\boldsymbol{X}^T_\lambda \boldsymbol{X}_\lambda)^{-1}\boldsymbol{X}^T_\lambda$.

Estudios previos (Eubank 1999) concluyen que una elección de λ que minimice el sesgo del estimador σ^2_λ no necesariamente conduce a una buena selección del estimador f_λ de f. Una estimación de la varianza σ^2 que no dependa de λ permitiría hacerla independiente del estimador de f.

En este contexto de los modelos no paramétricos se han propuesto varios estimadores de la varianza que no dependen de la elección de λ. La idea original fue propuesta por John Rice en 1984, cuyo estimador denotaremos σ^2_R y llamaremos *estimador de Rice*. Se define así:

$$\sigma^2_R = \frac{1}{2(n-1)}\sum_{i=2}^{n}(y_i - y_{i-1})^2 \tag{2.12}$$

Dada nuestra elección 2.2 de los valores de X, si $f \in C^1[0,1]$, entonces

$$\begin{aligned} y_i - y_{i-1} &= \epsilon_i - \epsilon_{i-1} + f(x_{i-1}) - f(x_i) \\ &= \epsilon_i - \epsilon_{i-1} + O(n^{-1}) \end{aligned} \tag{2.13}$$

Por lo que σ_R^2 depende en esencia de las contribuciones de los errores aleatorios que contienen toda la información acerca de σ^2.

Más adelante, Gasser, Sroka & Jennen-Steinmetz (1986) desarrollaron otra idea de Rice (1984), quien sugirió utilizar residuales obtenidos a partir de rectas ajustadas con tres puntos consecutivos para obtener otro estimador de la varianza. Para construir su estimador, Gasser y sus colegas definieron entonces unos *pseudo-residuales*, así:

$$\tilde{\epsilon_i} = \frac{(y_i - A_i y_{i-1} - B_i y_{i+1})}{(1 + A_i^2 + B_i^2)^{1/2}},\ i = 2, \ldots, n-1 \tag{2.14}$$

con

$$A_i = (x_{i+1} - x_i)/(x_{i+1} - x_{i-1})$$
$$B_i = (x_i - x_{i-1})/(x_{i+1} - x_{i-1})$$

El *estimador GSJS* se denotará σ_{GSJS}^2 y se define a partir de los pseudo-residuales 2.14 así:

$$\sigma_{GSJS}^2 = \frac{1}{n-2} \sum_{i=2}^{n-1} \tilde{\epsilon_i}^2 \tag{2.15}$$

El estimador GSJS busca resolver una limitación potencial del estimador de Rice en casos en los que se produzcan cambios muy fuertes de la función de regresión en subintervalos muy pequeños de X. La solución en este caso consiste en medir la diferencia entre cada y_i y la recta que une sus dos vecinos más cercanos y_{i-1} y y_{i+1} (Tong & Wang 2005) y usar estos pseudo-residuales para obtener un estimador insesgado de σ^2.

Luego, en 1990, Hall, Kay y Titterington (Hall, Kay & Titterington 1990) proponen otro estimador basado en diferencias sucesivas, siguiendo la misma idea de los estimadores de Rice y GSJS. Lo llamaremos *estimador HKT* y consecuentemente lo denotaremos σ_{HKT}^2.

Sean m_1 y m_2 dos enteros no negativos tales que $m_1 + m_2 = r$. El estimador σ_{HKT}^2 se basa en sucesiones de diferencias de orden r $\{d_k\}_{k=-m_1}^{m_2}$, sujetas a $\sum_{-m_1}^{m_2} d_k = 0$ y $\sum_{-m_1}^{m_2} d_k^2 = 1$.

El estimador σ_{HKT}^2 se define en general como:

$$\sigma_{HKT}^2 = \frac{1}{n-r} \sum_{l=m_1+1}^{n-m_2} \left(\sum_{k=-m_1}^{m_2} d_k y_{k+l} \right)^2 \tag{2.16}$$

En el caso $m_1 = 0$ y $m_2 = r$, el estimador HKT sería:

$$\sigma^2_{HKT_r} = \frac{1}{n-r}^{-1} \sum_{l=1}^{n-r} \left(\sum_{k=0}^{r} d_k y_{k+l} \right)^2 \tag{2.17}$$

El estimador HKT es un estimador insesgado de σ^2 que generaliza los dos estimadores previos de Rice y GSJS. El estimador de Rice coincide con el estimador HKT cuando $r = 1$ con $m_1 = 0$ y $m_2 = 1$. Y el estimador GSJS se obtiene a partir del estimador HKT con $m_1 = 0$ y $m_2 = 2$, si el diseño es equidistante, como se asume en este capítulo.

Estos estimadores están pensados bajo el modelo homocedástico 2.1. En los casos en los cuales el modelo admite una función de varianza, por ejemplo en un modelo del tipo $Y_i = f(x_i) + \sqrt{V(x_i)}\epsilon_i$, $i = 1, \ldots, n$, se han propuesto nuevos estimadores de la varianza, algunos de los cuales se pueden consultar en Levine (2006) y Brown & Levine (2007).

2.2.2. Algunas ideas sobre cómo conducir inferencias

Tal como en los modelos lineales, no parece razonable conservar una función f_j que tenga asociado un $\beta_j = 0$. Así que una primera prueba formal del análisis de regresión será contrastar las hipótesis

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

Un posible estadístico de prueba, adaptado de los modelos lineales, sería:

$$Z_j = \beta_{\lambda j} / [(\boldsymbol{X}_\lambda^T \boldsymbol{X}_\lambda)_j \sigma^2]^{1/2} \tag{2.18}$$

donde $(\boldsymbol{X}_\lambda^T \boldsymbol{X}_\lambda)_j$ es el elemento j de la diagonal principal de la matriz $\boldsymbol{X}_\lambda^T \boldsymbol{X}_\lambda$. En lugar de σ^2 podríamos utilizar uno de los estimadores propuestos en la Sección 2.2.1.

De igual forma, un intervalo de confianza para f sería:

$$\boldsymbol{f}_\lambda(x) \pm Z_{\alpha/2} [\sigma^2 \boldsymbol{f}_\lambda(x)^T (\boldsymbol{X}_\lambda^T \boldsymbol{X}_\lambda)^{-1} \boldsymbol{f}_\lambda(x)]^{1/2} \tag{2.19}$$

con $\boldsymbol{f}_\lambda(x) = (f_1(x), \ldots, f_\lambda(x))^T$

El análisis de homocedasticidad y demás pruebas de diagnóstico se podrían tomar prestados de los métodos clásicos de análisis de regresión.

Sin embargo, estas ideas requieren de evaluar la calidad del estimador, por lo que una pregunta natural es si el estimador f_λ definido en 2.8 es o no un estimador consistente de f.

2.2.3. ¿Es f_λ un estimador consistente de f?

Los métodos estadísticos tienen una incertidumbre inherente: la estimación obtenida a partir de una muestra no coincidirá necesariamente con lo que se busca estimar en la población.

Si el problema que nos ocupara fuera estimar un parámetro de una población y dispusiéramos de un estimador insesgado, esta incertidumbre se expresaría en términos de la varianza de este estimador. Sería además deseable que el estimador convergiera al parámetro. Pero que converja no es suficiente: es mejor si converge a la mayor velocidad posible. En la estimación paramétrica se expresa la incertidumbre inherente al método en términos de la varianza del estimador, que usualmente converge a cero a una velocidad igual al inverso del tamaño de la muestra en la cual se basa la estimación $(1/n)$.

En los métodos de regresión no paramétrica los estimadores en general no son insesgados, por lo que la varianza del estimador no será suficiente para evaluar la incertidumbre inherente a estos métodos. En consecuencia, preferiremos un estimador que tenga el menor sesgo posible para n grande para lo cual acudiremos a una de las propiedades de los estimadores cercana a este propósito que es la consistencia. Y luego acudiremos a la eficiencia del estimador, una propiedad asociada con la tasa de convergencia que nos permitirá elegir entre estimadores consistentes.

Para continuar con nuestro propósito de estudiar la consistencia y la eficiencia de los estimadores, debemos introducir algunas definiciones útiles en el contexto de los modelos de regresión no paramétrica.

En primer lugar, definiremos la *pérdida* de f_λ para estimar f:

$$L_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - f_\lambda(x_i))^2 \tag{2.20}$$

Y definiremos el *riesgo* de f_λ para estimar f como el valor esperado de la pérdida:

$$R_n(\lambda) = E[L_n(\lambda)] \tag{2.21}$$

Comentario: Algunos autores (e.g. Hastie & Tibshirani (1990)) prefieren llamar *error cuadrático medio*, denotado $MSE(\lambda)$, a lo que aquí definimos como riesgo. Mantendremos este último nombre, sugerido por Eubank (1999).

Si nos detenemos por un momento en las expresiones 2.20 y 2.21 y tratamos de asociar estas definiciones con resultados similares en los métodos clásicos, encontraremos que el riesgo es similar al error cuadrático medio de un estimador $\hat{\theta}$ de un parámetro θ (que se desea estimar). En el caso de nuestra definición de riesgo, tenemos un estimador f_λ que deseamos utilizar para estimar una función de regresión f. Una diferencia crucial es que $\hat{\theta}$ y θ están definidos en $\Re$, mientras que f_λ y f están definidos en $L_2[0,1]$. Podríamos decir, con las debidas precauciones, que el riesgo podría verse como una generalización del error cuadrático medio.

El riesgo del estimador (2.8) es:

$$R_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\{f_\lambda(x_i) - E[f_\lambda(x_i)]\}^2 + E\{f_\lambda(x_i) - E[f_\lambda(x_i)]\}^2 \quad (2.22)$$

En la expresión del riesgo de f_λ, el primer término de 2.22 representa el promedio de los sesgos al cuadrado, mientras que el segundo término representa la varianza del estimador f_λ. Nótese que este resultado es similar al del error cuadrático medio en el caso paramétrico.

Para evaluar la convergencia del primer término, que recoge los sesgos al cuadrado, tenemos que entre más términos tenga el estimador, más se parecerá a la función. Es decir, a medida que λ es más grande, la aproximación de f_λ a f es cada vez mejor y por tanto este promedio crecerá lentamente a medida que λ crece. Es decir, necesitaremos que $\lambda \to \infty$ a medida que $n \to \infty$. En realidad una mayor precisión de esta tendencia requiere hacer más explícitas las funciones f_i, por lo que diremos por ahora que este resultado es muy intuitivo.

Por otra parte, $Var(f_\lambda) = E\{f_\lambda(x_i) - E[f_\lambda(x_i)]\}^2 = \lambda\sigma^2/n$. Es decir, la varianza del estimador f_λ es $\lambda\sigma^2/n$, que crece a medida que λ crece, lo que significa que para que el riesgo de f_λ decaiga a cero se requiere que $\lambda/n \to 0$. O sea que para que f_λ sea consistente se requiere que λ crezca más lentamente que n.

Acordemos por lo pronto que nos satisface la consistencia de f_λ como estimador de f y que nuestras inferencias de la Sección

2.2.2 son aplicables. Quedaría entonces por resolver una de nuestras preguntas abiertas: ¿Cómo elegir λ? Sin embargo, antes de enfrentar este problema discutiremos un ejemplo de un estimador particular del tipo 2.8.

2.2.4. Un ejemplo

Los datos representados en la Figura 2.1 corresponden a mediciones del promedio horario de Monóxido de Carbono (CO) en una estación ubicada en el centro de la ciudad de Cali, Colombia (Calle 15), el 1 de marzo de 2004. Los datos provienen de la red de monitoreo de la calidad del aire de Cali, administrada por el Departamento Administrativo para la Gestión del Medio Ambiente (DAGMA).

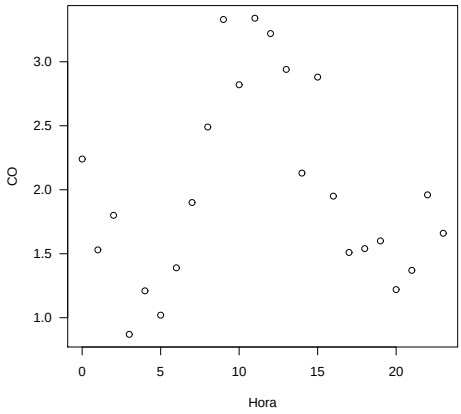


Figura 2.1: Promedios horarios de CO en el centro de Cali el 1 de marzo de 2004

Nuestro propósito es ajustar un modelo del tipo 2.1 a este conjunto de observaciones de un día.

Supongamos que disponemos de la siguiente sucesión ortonormal completa (CONS) de funciones:

$$\begin{aligned} f_1(x) &= 1 \\ f_j(x) &= \sqrt{2}cos((j-1)\pi x),\ j = 2, \ldots \end{aligned} \tag{2.23}$$

Usando la CONS 2.23, nuestro estimador 2.8 tomaría la siguiente forma:

$$f_\lambda(x) = \sum_{j=1}^{\lambda} \beta_{\lambda j}\sqrt{2}cos((j-1)\pi x) \tag{2.24}$$

Podemos estimar los $\beta_{\lambda j}$ utilizando el método de mínimos cuadrados tal como se describió en 2.7, en la que la matriz $\boldsymbol{X}_\lambda$ será:

$$\begin{aligned}\boldsymbol{X}_\lambda &= [\{f_j(x_i)\}_{j=1,\ldots,\lambda;\ i=1,\ldots,n}] \\ &= [1|\{f_j(x_i)\}_{j=2,\ldots,\lambda;\ i=1,\ldots,n}] \\ &= [1|\sqrt{2}cos((j-1)\pi x),\ j=2,\ldots,\lambda;\ i=1,\ldots,n]\end{aligned} \tag{2.25}$$

Supongamos que para los datos de la figura 2.1 decidimos estimar f usando $\lambda = 4$. Es decir, lo que nos proponemos es utilizar las primeras cuatro funciones de la CONS de cosenos que se muestran en la figura 2.2, ponderarlas utilizando las estimaciones $\beta_{4j}, j = 1, 2, 3, 4$ y estimar f_4 así:

$$f_4(x) = \beta_{41}f_1 + \beta_{42}f_2 + \beta_{43}f_3 + \beta_{44}f_4 \tag{2.26}$$

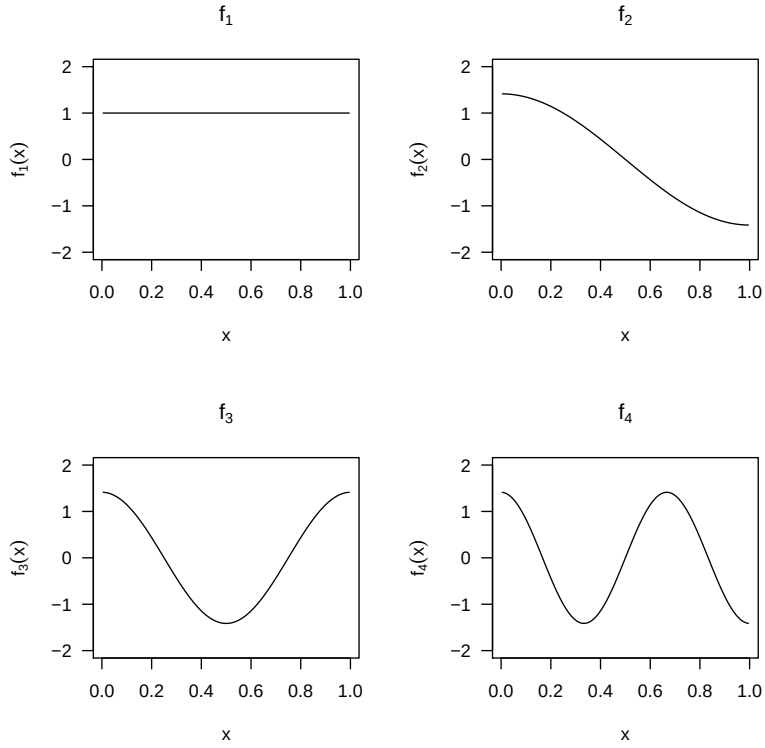


Figura 2.2: Primeras cuatro funciones de la CONS de cosenos

Los datos se presentan en la tabla 2.1. Los valores de X corresponden a las horas del día (variable HORA) y los valores de Y a las mediciones del promedio horario de Monóxido de Carbono (CO)

en mg/m^3. Sin pérdida de generalidad, los valores de X se convertirán a valores entre 0 y 1.

Tabla 2.1: Concentraciones promedio horarias (mg/m) de Monóxido de Carbono (CO) en el centro de Cali

HORA	CO	HORA	CO
0	2.24	12	3.22
1	1.53	13	2.94
2	1.80	14	2.13
3	0.87	15	2.88
4	1.21	16	1.95
5	1.02	17	1.51
6	1.39	18	1.54
7	1.90	19	1.60
8	2.49	20	1.22
9	3.33	21	1.37
10	2.82	22	1.96
11	3.34	23	1.66

Tenemos en este caso que $\beta_4 = (2.0281, -0.0232, -0.5329, 0.0117)$. El estimador f_4 de la función de regresión f se representa en la figura (2.3), que parece indicar que $\lambda = 4$ no es una muy buena elección.

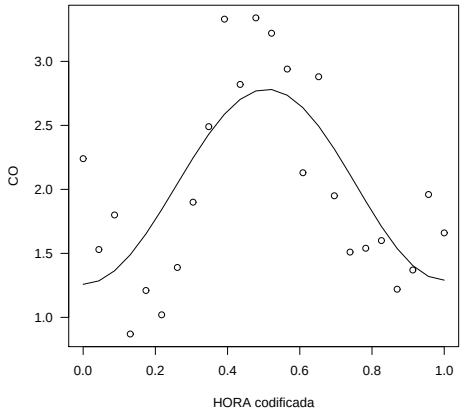


Figura 2.3: Estimación del comportamiento de la contaminación por CO en el centro de Cali el 1 de marzo de 2004, usando $\lambda = 4$ con series de cosenos

Pero, entonces, ¿Cuál valor de λ sería una "buena elección"? Hasta este momento no tenemos una respuesta. Para tratar de ganar un poco más de conocimiento sobre esta elección, la figura 2.4 ilustra la forma que tomaría el estimador de la función de regresión usando cuatro valores de λ: 1 y 24 (los valores mínimo y máximo posibles para λ)

y 6 y 9. En el primer caso, un valor muy pequeño de λ conducirá a que el estimador de la función de regresión sea el promedio de Y. Así que en tal caso se minimiza la varianza del estimador, aunque se maximiza el sesgo. Si λ es muy grande, el estimador reproducirá las observaciones de tal manera que este estimador será insesgado, pero de máxima varianza. Si λ toma el valor 6, el estimador parece seguir bien el comportamiento de las observaciones, aunque con λ igual a 9 se tiene una mejor representación hacia la media noche. Los estimadores f_6 y f_9 tendrán combinaciones de sesgo y varianza cuya suma preferiríamos que fuera los más pequeña posible. Es decir, un buen valor de λ podría ser aquel que minimizara el riesgo $R_n(\lambda)$ del estimador.

Este ejemplo nos permite entonces pasar a discutir el problema de la elección de λ.

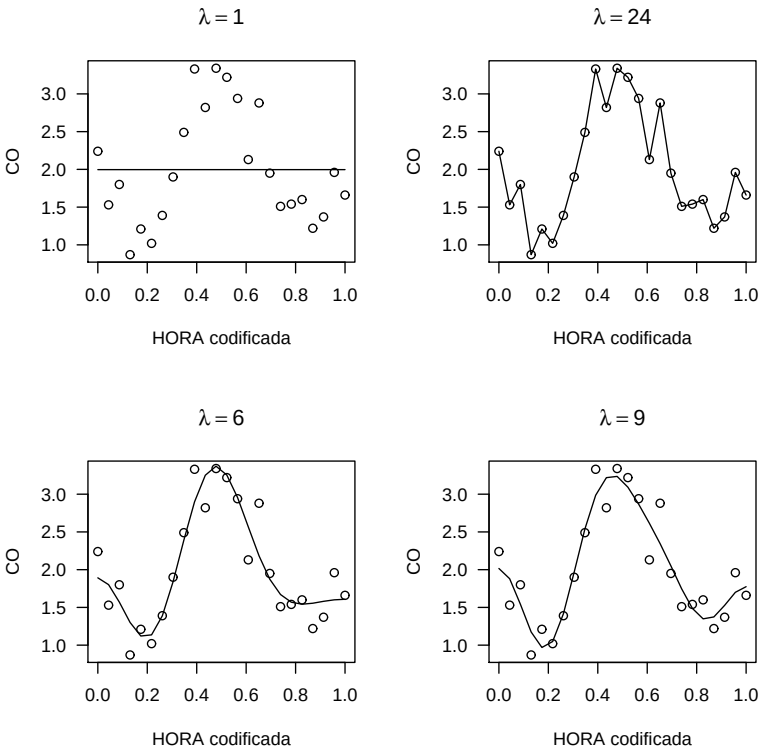


Figura 2.4: Estimación del comportamiento de la contaminación por CO en el centro de Cali el 1 de marzo de 2004, usando $\lambda = 1, 24, 6$ y 9. Con $\lambda = 1$ el estimador coincide con el promedio de las concentraciones de CO; con $\lambda = 24$ el estimador reproduce los datos; con $\lambda = 6$ el estimador parece seguir mejor el comportamiento de los datos, al parecer mejorando aún más con $\lambda = 9$, especialmente hacia final del día.

2.2.5. ¿Cómo se elige el valor λ?

En la sección 2.2.3 definimos dos propiedades de un estimador: la pérdida $L(\lambda)$ y el riesgo $R(\lambda)$, que permiten medir el desempeño de un estimador de tal manera que valores pequeños de $L(\lambda)$ y $R(\lambda)$ se asocian con buenos estimadores. Esta noción de "bondad" de los estimadores se construye desde la perspectiva de la minimización de la suma de cuadrados de los residuales, aunque diversas personas podrían proponer otras formas de decidir si un estimador es o no "bueno". Adoptaremos la idea de elegir los estimadores con criterios de calidad similares a los de otras técnicas de análisis de regresión, que se basan en general en esta suma de cuadrados de los residuales.

Si miramos la definición de la pérdida 2.20, esta se refiere a una muestra particular de tamaño n, por lo que el valor de λ que minimice $L(\lambda)$ nos permitirá construir el mejor estimador f_λ de f basado en esta muestra en particular. Este λ es óptimo para este conjunto de observaciones. Entre tanto, la definición del riesgo (2.21) permite ampliar la selección de un óptimo al caso de muestreo repetido.

Si uno intenta minimizar $L(\lambda)$ y $R(\lambda)$, se encuentra con la dificultad que ambos dependen de la función de regresión f que se desea estimar. Esto nos lleva al uso de la suma de cuadrados de las diferencias entre las observaciones y_i y las estimaciones $f_\lambda(x_i)$, que denotaremos $RSS(\lambda)$.

$$RSS(\lambda) = \sum_{i=1}^{n} (y_i - f_\lambda(x_i))^2 \tag{2.27}$$

En lo sucesivo simplificaremos la notación así:

$$\begin{aligned} f_i &= f(x_i), \quad i = 1, \ldots, n \\ f_{\lambda i} &= f_\lambda(x_i), \quad i = 1, \ldots, n \end{aligned}$$

Denotaremos además $\boldsymbol{f}$ al vector de valores de la función f y $\boldsymbol{f_\lambda}$ al vector de valores del estimador f_λ, en los puntos de diseño x_i. $\boldsymbol{y}$ será el vector de respuestas.

Entonces la suma de cuadrados de los residuales se puede reescribir como:

$$\begin{aligned} RSS(\lambda) &= (\boldsymbol{y} - \boldsymbol{f_\lambda})^T (\boldsymbol{y} - \boldsymbol{f_\lambda}) \\ &= \boldsymbol{y^T (I - S_\lambda) y} \end{aligned}$$

donde $\boldsymbol{S_\lambda} = \boldsymbol{(X_\lambda^T X_\lambda)^{-1} X_\lambda^T X_\lambda}$, tal como lo definimos en la expresión (2.11).

De nuestra definición 2.21 sabemos que (Graybill 2001):

$$\begin{aligned} R(\lambda) &= \frac{1}{n}\sum_{i=1}^{n} E(\boldsymbol{f_i} - \boldsymbol{f_{\lambda i}})^2 \\ &= \frac{1}{n} E[(\boldsymbol{f} - \boldsymbol{f_\lambda})^{\boldsymbol{T}}(\boldsymbol{f} - \boldsymbol{f_\lambda})] \\ &= \frac{1}{n}\boldsymbol{f^T}(\boldsymbol{I} - \boldsymbol{S_\lambda})^{\boldsymbol{2}}\boldsymbol{f} + \frac{1}{n}\sigma^2 \mathrm{tr}[\boldsymbol{S_\lambda^2}] \end{aligned} \tag{2.28}$$

Por otra parte,

$$\begin{aligned} E[RSS(\lambda)] &= \boldsymbol{f^T}(\boldsymbol{I} - \boldsymbol{S_\lambda})^{\boldsymbol{2}}\boldsymbol{f} + \sigma^2 \mathrm{tr}[(\boldsymbol{I} - \boldsymbol{S_\lambda})^{\boldsymbol{2}}] \\ &= \boldsymbol{f^T}(\boldsymbol{I} - \boldsymbol{S_\lambda})^{\boldsymbol{2}}\boldsymbol{f} + \sigma^2 \mathrm{tr}[\boldsymbol{S_\lambda^2}] + n\sigma^2 - 2\sigma^2 \mathrm{tr}[\boldsymbol{S_\lambda}] \end{aligned} \tag{2.29}$$

Si insertamos 2.28 en 2.29, tenemos que:

$$E[RSS(\lambda)] = nR(\lambda) + n\sigma^2 - 2\sigma^2 \mathrm{tr}[\boldsymbol{S_\lambda}] \tag{2.30}$$

De donde se sigue que $\frac{1}{n}RSS$ es un estimador sesgado de $R(\lambda)$ con sesgo igual a $\sigma^2 - \frac{2}{n}\sigma^2 \mathrm{tr}[\boldsymbol{S_\lambda}]$. En consecuencia, un estimador insesgado del riesgo $R(\lambda)$ del estimador f_λ sería:

$$\hat{R}_U(\lambda) = \frac{1}{n}RSS(\lambda) - \sigma^2 + \frac{2}{n}\sigma^2 \mathrm{tr}[\boldsymbol{S_\lambda}] \tag{2.31}$$

Nuestro estimador insesgado $\hat{R}_U(\lambda)$ tropieza con la dificultad de que no conocemos σ^2. Una solución sería utilizar uno de los estimadores de σ^2 que presentamos en la sección 2.11. Si adoptamos la notación genérica $\hat{\sigma}^2$ para nuestro estimador de la varianza que no depende de la elección de λ (sección 2.11), entonces un posible estimador de $R(\lambda)$, atribuido a Wahba (1990), sería:

$$\hat{R}(\lambda) = \frac{1}{n}RSS(\lambda) + \frac{2}{n}\hat{\sigma}^2 \mathrm{tr}[\boldsymbol{S_\lambda}] - \hat{\sigma}^2 \tag{2.32}$$

Este estimador se conoce en la literatura y en los ambientes computacionales como estimador $UBRE$ por las siglas en inglés de *estimador insesgado del riesgo*.

Utilicemos este estimador para identificar el valor de λ más adecuado para la estimación de f mediante series de cosenos en nuestro ejemplo de CO en la Calle 15 de Cali. Sabemos que $\lambda \in \{1, 2, \ldots, 24\}$, el número de funciones f_i que conservaremos en la estimación de

f. Calcularemos para cada valor posible de λ el estimador $UBRE$ del riesgo $R(\lambda)$, al cual denotaremos $\hat{R}(\lambda)$. Para la estimación de la varianza usaremos el estimador HKT. Los resultados se despliegan en la tabla 2.2 y se representan en la figura 2.5.

Tabla 2.2: Valores estimados del riesgo usando el estimador $UBRE$ ($\hat{R}(\lambda)$) para cada valor posible de λ, en el problema de contaminación por CO en la Calle 15 de Cali

λ	$\hat{R}(\lambda)$	λ	$\hat{R}(\lambda)$
1	0.429	13	0.065
2	0.439	14	0.071
3	0.154	15	0.074
4	0.164	16	0.073
5	0.025	17	0.082
6	0.017	18	0.092
7	0.013	19	0.102
8	0.022	20	0.111
9	0.032	21	0.121
10	0.042	22	0.105
11	0.052	23	0.110
12	0.062	24	0.120

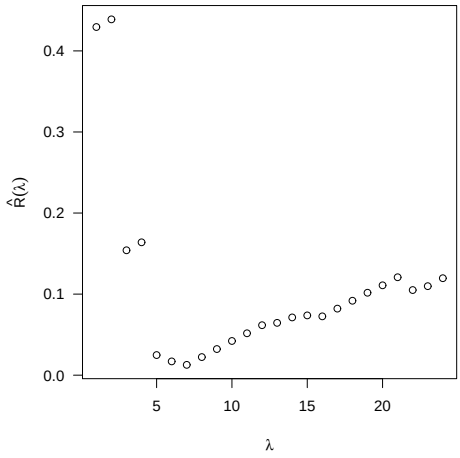


Figura 2.5: El mínimo de $\hat{R}(\lambda)$ se presenta cuando $\lambda = 7$.

De acuerdo con los resultados de la tabla 2.2 y de la figura 2.5, el valor óptimo de λ, de acuerdo con el estimador $UBRE$, es $\lambda = 7$. Es decir, basados en este indicador, elegiremos a f_7 como el mejor estimador de f en el problema de la contaminación por CO en la Calle 15 de Cali, usando el estimador de cosenos.

El estimador $UBRE$ es una opción para elegir λ que podría considerarse como el primer intento sustentado de selección. Su principal limitación es su dependencia del conocimiento de la varianza σ^2, por lo que sería interesante indagar si es posible proponer métodos de selección de λ que no dependan de ese conocimiento.

Para lograrlo, definamos primero otra medida de calidad del estimador f_λ. Supongamos que nos proponemos obtener n nuevas observaciones de Y que se supone pueden ser modelados de la misma forma que los datos originales. Llamaremos $\boldsymbol{y_N}$ al vector de nuevas observaciones. Asumamos que se satisface que:

$$y = f(x) + \epsilon,\ i = 1, \ldots, n \tag{2.33}$$

con f la misma función de regresión del modelo 2.1 y los ϵ variables aleatorias no correlacionadas entre sí ni con los ϵ, con varianza común σ^2.

Digamos que queremos usar nuestro estimador f_λ para predecir los y. Definamos entonces el *riesgo de predicción* $P(\lambda)$ como sigue:

$$P(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E[y - f]^2 \tag{2.34}$$

Nuestra definición de las nuevas observaciones nos permite verificar que el riesgo de predicción $P(\lambda)$ y el riesgo $R(\lambda)$ se relacionan de la siguiente manera:

$$P(\lambda) = \sigma^2 + R(\lambda) \tag{2.35}$$

lo que permitiría definir un estimador no insesgado de $P(\lambda)$ basado en el estimador $UBRE$ definido en (2.32). Pero los resultados serían los mismos ya que estos dos métodos son equivalentes en el sentido que el λ que hace mínimo $P(\lambda)$ es el mismo que minimiza $R(\lambda)$.

Si uno dispusiera de un segundo conjunto de n datos, estimaría f_λ con los primeros n datos y buscaría el λ que minimice $P(\lambda)$ usando el segundo conjunto de datos. En la práctica, sin embargo, sería mejor estimar f utilizando los $2n$ datos, lo que nos lleva a continuar nuestra búsqueda de un nuevo procedimiento para elegir λ que no dependa de σ^2.

Una opción es dividir el conjunto de n observaciones en n sub-muestras de tamaño $n - 1$ mediante el mecanismo de dejar por fuera una observación diferente cada vez. Si denotamos $f_{\lambda(i)}$ a la estimación de f_i obtenida al suprimir de la muestra la

observación i, entonces la observación y_i sería una observación adicional que podríamos utilizar para construir un estimador de $P(\lambda)$ que denotaremos $CV(\lambda)$ y que llamaremos criterio de *validación cruzada*:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_{\lambda(i)})^2 \tag{2.36}$$

Calcular el criterio CV por el método propuesto en (2.36) es complejo, porque requiere una gran carga de procesamiento computacional. Green & Silverman (2000) sugieren simplificar este cálculo utilizando en cambio:

$$CV_{GS}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - f_{\lambda i}}{1 - S_{\lambda ii}}\right)^2 \tag{2.37}$$

donde $S_{\lambda ii}$ es el elemento i de la matriz S_λ.

Otra posible solución se debe a Wahba (1990), quien sugiere utilizar otro criterio llamado *validación cruzada generalizada* $GCV(\lambda)$, definido por Green & Silverman (2000) como

$$GCV_{GS}(\lambda) = n^{-1}\frac{\sum_{i=1}^{n}(y_i - f_{\lambda i})^2}{(1 - n^{-1}\mathrm{tr}[S_\lambda])^2} \tag{2.38}$$

Comparando las expresiones (2.36) y (2.37) se ve que los residuales obtenidos al eliminar una de las observaciones se pueden obtener a partir de los residuales mediante el mecanismo de dividirlos entre el factor $1 - S_{\lambda ii}$. La idea del criterio GCV es reemplazar estos factores por su promedio $1 - n^{-1}\mathrm{tr}[S_\lambda]$. Como en el numerador aparece la suma de cuadrados de los residuales RSS, una expresión más sencilla sería (Eubank 1999):

$$GCV(\lambda) = \frac{n^{-1}RSS(\lambda)}{(n^{-1}\mathrm{tr}[I - S_\lambda])^2} \tag{2.39}$$

Aunque la palabra "generalizada" deja la impresión de que el segundo criterio generaliza el primero, esto no es en general cierto y se trata de criterios diferentes que permiten estimar el riesgo de predicción $P(\lambda)$. Wahba (1990) justifica el uso de este criterio como un buen método de selección de λ demostrando que si $n^{-1}\mathrm{tr}[S_\lambda] < 1$, entonces la diferencia entre $E[GCV(\lambda)]$ y $P(\lambda)$ relativa al tamaño de $R(\lambda)$ será pequeña, especialmente para tamaños de muestra grande.

Veamos cómo operan los criterios $CV(\lambda)$ y $GCV(\lambda)$ en nuestro ejemplo de la contaminación del aire debida al CO en la Calle 15 de

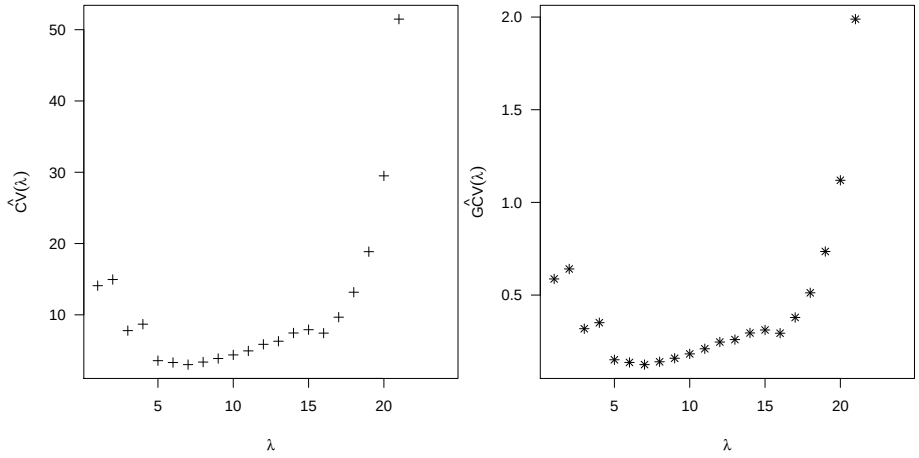


Figura 2.6: El mínimo de $\hat{CV}(\lambda)$ y $G\hat{CV}(\lambda)$ se presenta cuando $\lambda = 7$, al igual que con $\hat{R}(\lambda)$.

Cali. Ambos criterios conducen a elegir el valor $\lambda = 7$ como el más adecuado para estimar f en el ejemplo de la estimación de la función de regresión con los datos de contaminación del aire por CO en la calle 15 de Cali, cuando se usa una serie de cosenos. La figura 2.7 se representa la curva ajustada a los datos usando $\lambda = 7$.

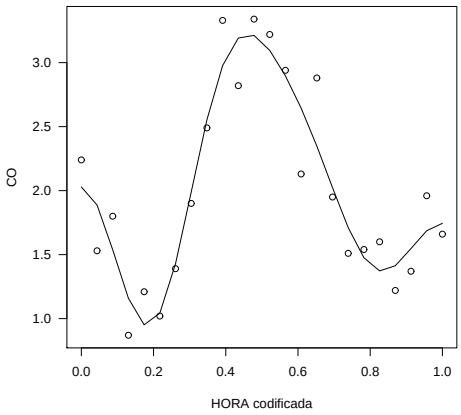


Figura 2.7: Estimación con series de cosenos del comportamiento diario de los promedios horarios de CO en el centro de Cali el 1 de marzo de 2004, usando el λ óptimo ($\lambda = 7$) para este conjunto de datos

2.3. Importancia de los estimadores de series

Aunque en la práctica no se utilizan los estimadores de series, su importancia radica, por una parte, en que permiten visualizar el

funcionamiento de la técnica. Pero más que esto, porque estudiando sus propiedades es posible expresarlos en forma tal que se convierten en una base teórica para la construcción de estimadores más eficientes que funcionan de manera similar.

En primer lugar, la serie de cosenos es una CONS. Para ver sus propiedades de ortonormalidad, consideremos los siguientes dos resultados. El primero es:

$$n^{-1}\sum_{i=1}^{n} cos(j\pi x_i)cos(k\pi x_i) = \frac{1}{2}\delta_{jk}, \quad j,k = 1,\ldots,n-1 \tag{2.40}$$

donde δ_{jk} es la función delta de Kronecker:

$$\delta_{jk} = \begin{cases} 1, & \text{si } j = k \\ 0, & \text{en otro caso} \end{cases}$$

Y el segundo resultado es:

$$n^{-1}\sum_{i=1}^{n} cos(j\pi x_i) = 0, \quad j = 1,\ldots,n-1 \tag{2.41}$$

Una implicación de estos dos resultados es que $\boldsymbol{X_\lambda^T X_\lambda} = n\boldsymbol{I}$, de donde los coeficientes de nuestro estimador de series de cosenos 2.24 pueden reescribirse ahora como:

$$\begin{aligned} \beta_{\lambda 1} &= n^{-1}\sum_{i=1}^{n} y_i = \bar{y} \\ \beta_{\lambda j} &= n^{-1}\sum_{i=1}^{n} y_i\sqrt{2}cos((j-1)\pi x_i), \quad j = 2,\ldots,\lambda \end{aligned} \tag{2.42}$$

Con estos resultados, nuestro estimador f_λ de la función de regresión que definimos en 2.24 tomaría la siguiente forma:

$$f_\lambda(x) = \bar{y} + \sum_{j=2}^{\lambda} \beta_{\lambda j}\sqrt{2}cos((j-1)\pi x) \tag{2.43}$$

Lo más interesante del resultado 2.43 es que puede expresarse de la siguiente nueva forma, que nos servirá para ilustrar nuevos estimadores en lo que sigue.

$$f_\lambda(x) = n^{-1}\sum_{i=1}^{n} y_i K_\lambda(x, x_i) \tag{2.44}$$

donde:

$$K_\lambda(x,s) = 1 + cos\left(\frac{\lambda\pi(x-s)}{2}\right) D\left(\frac{x-s}{2}; \lambda - 1\right) + cos\left(\frac{\lambda\pi(x+s)}{2}\right) D\left(\frac{x+s}{2}; \lambda - 1\right) \quad (2.45)$$

En 2.45 la función D se conoce como kernel Dirichlet que proviene del análisis de series de Fourier y se define como:

$$D(u,k) = \frac{sen(k\pi u)}{sen(\pi u)}$$

Para obtener la expresión 2.45 se requiere disponer de dos resultados previos. Primero, que:

$$cos(x) + cos(y) = 2cos\left(\frac{1}{2}(x+y)\right) cos\left(\frac{1}{2}(x-y)\right)$$

y segundo, que:

$$\sum_{j=1}^{k} cos(j\pi x) = cos\left(\frac{(k+1)\pi x}{2}\right) D\left(\frac{x}{2}; k\right) \quad (2.46)$$

El resultado 2.46 se debe a Gradshtein & Ryzhik (1980), citado por Eubank (1999, Pág. 88).

Usando 2.41 se encuentra que $\sum_{i=1}^{n} K_\lambda(x, x_i) = n$. Y aunque esta sumatoria incluye términos negativos, si uno acepta el uso de ponderadores negativos, entonces 2.44 es un promedio ponderado de las observaciones y_i.

La figura 2.8 ilustra el comportamiento de la función ponderadora 2.45 para $x = 0.5$ y $\lambda = 10$. Se observa que para la estimación de la función de regresión, esta función ponderadora asigna los pesos más altos a las observaciones y_i asociadas con valores de x_i cercanos a $x = 0.5$. Funciones ponderadoras que usan esta idea como principio y que no generan pesos negativos son el sustento para la suavización kernel que discutiremos más adelante.

2.4. El modelo de regresión polinómica

Asumamos por un momento que disponemos de una función de regresión f conocida. De acuerdo con el Teorema de Taylor, si tenemos

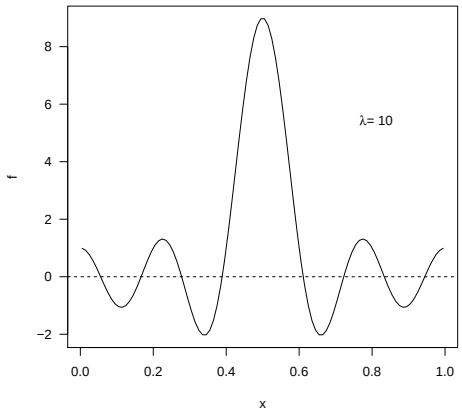


Figura 2.8: Ponderador Kernel Dirichlet para $\lambda = 10$.

una función continua f en el intervalo cerrado $[a, b]$ que tiene $\lambda + 1$ derivadas continuas en el intervalo (a, b), entonces si x y c son dos puntos en (a, b), se sigue que $f(x)$ se puede representar como:

$$f(x) = f(c) + f'(c)(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \ldots + \frac{f^{(\lambda)}(c)}{\lambda!}(x-c)^{\lambda} + r_{\lambda}(\xi) \tag{2.47}$$

donde el término $r_\lambda(\xi)$ depende de la derivada $f^{(\lambda+1)}$ de f para cierto ξ entre x y c, y se conoce como el λ-*residuo*.

La expresión de f usando el Teorema de Taylor 2.47 puede reescribirse como:

$$f(x) = \sum_{j=0}^{\lambda} \theta_j f_j(x) + r_{\lambda}(x) \tag{2.48}$$

donde en la expresión 2.48 se tiene que $\theta_1 = f(c)$, $\theta_j = \frac{f^{(j-1)}(c)}{(j-1)!}, j = 2, \ldots, \lambda$ y $f_j(x) = (x-c)^{j-1}, j = 1, \ldots, \lambda$.

Es decir, la función f se puede representar como la suma un polinomio $\pi_\lambda(x)$ de grado λ más un residuo $r_\lambda(x)$. Hay varias propuestas sobre cómo calcular $r_\lambda(x)$ que no se incluyen aquí, pero que requieren todas la existencia de $f^{(\lambda+1)}(x)$. Ahora bien, que en la medida que $r_\lambda(x)$ sea más *pequeño*, entonces $\pi_\lambda(x)$ se parecerá más a $f(x)$. En resumen, existe un λ tal que:

$$f(x) \doteq \sum_{j=0}^{\lambda} \theta_j f_j(x) \tag{2.49}$$

Supongamos ahora que hemos colectado información sobre dos variables X y Y que se relacionan según el modelo 2.1, por lo que la función de regresión no es conocida. Si asumimos que f es una función de regresión que tiene $\lambda + 1$ derivadas continuas y puede representarse en la forma 2.48, entonces, si $r_\lambda(x)$ es uniformemente pequeño, podríamos escribir:

$$f(x) \doteq \sum_{j=0}^{\lambda} \theta_j f_j(x) + \epsilon_j \tag{2.50}$$

Diremos entonces que nuestras observaciones se comportan aproximadamente según un *modelo de regresión (no-paramétrica) polinómica*, el cual presupone que los residuos $r_\lambda(x_1), \ldots, r_\lambda(x_n)$ de la expansión de f en series de Taylor 2.48 han sido incorporados en los errores aleatorios del modelo 2.50. Para encontrar los coeficientes θ_j se podría utilizar el método de mínimos cuadrados, o cualquiera otro. Pero su estimación no es un propósito de la regresión polinómica.

Por otra parte, si f se puede aproximar bien utilizando el polinomio $\pi_\lambda(x)$ y si los residuos en los x_i son pequeños en comparación con los errores ϵ_i, entonces el modelo 2.50 debería trabajar bien. En otro caso, podríamos estar en problemas con este modelo, porque como tanto los residuos como los errores con desconocidos, no hay forma de saber si esta aproximación basada en el Teorema de Taylor es aplicable para un conjunto de datos y un λ dados. Nuevos modelos que discutiremos más adelante en este texto superan estas limitaciones de la regresión polinómica.

En resumen, la regresión polinómica se basa en el Teorema de Taylor para encontrar un polinomio $\pi_\lambda(x)$ que aproxima la función de regresión desconocida f. Y por lo tanto requiere que f sea continua y tenga por lo menos $\lambda + 1$ derivadas continuas.

Anexo 1

El espacio $L_2[0, 1]$

El espacio $L_2[0, 1]$ se define como el espacio funcional de todas las funciones cuadrado integrables en $[0, 1]$. Sean f_1 y f_2 dos funciones en $L_2[0, 1]$. Entonces la norma de f_i se define como

$$\|f_i\| = \left\{ \int_0^1 f_i^2(x)dx \right\}^{1/2}, i = 1, 2 \tag{2.51}$$

mientras que el producto interno de f_1 y f_2 es

$$< f_1, f_2 >= \int_0^1 f_1(x) f_2(x) dx \tag{2.52}$$

Sucesiones ortonormales completas *CONS*

En un espacio dimensional finito, por ejemplo en $\Re^p$, siempre es posible representar cualquier elemento del espacio usando una combinación lineal de elementos en una base del espacio. Pero el espacio $L_2[0,1]$ es un espacio de dimensión infinita por lo que las siguientes definiciones se requieren para poder concluir que en efecto es posible representar cualquier elemento de $L_2[0,1]$ a partir de un conjunto de elementos que formen una base de $L_2[0,1]$.

1. Dos funciones $f_1, f_2 \in L_2[0,1]$ se dicen ortogonales si $< f_1, f_2 >= 0$. La ortogonalidad de f_1 y f_2 se denota $f_1 \perp f_2$.

2. Una sucesión de funciones $\{f_j\}_{j=1}^{\infty} \in L_2[0,1]$ se dice ortonormal si las f_j son ortogonales por parejas y $\|f_j\| = 1$ para todo j.

3. Una sucesión de funciones $\{f_j\}_{j=1}^{\infty} \in L_2[0,1]$ se dice que es una *sucesión ortonormal completa* (*CONS* por sus iniciales en inglés) si $f \perp f_j$ para todo j implica que $f = 0$.

Las definiciones de ortogonalidad y ortonormalidad en $L_2[0,1]$ son similares a las mismas definiciones en $\Re^p$, mientras que la definición de una *CONS* implica que la única función ortogonal a todas las funciones f_i es la función cero.

Coeficientes de Fourier

Necesitamos ahora garantizar que en efecto una función f de $L_2[0,1]$ puede representarse como una combinación lineal de una colección de funciones $\{f_j\}_{j=1}^{\infty}$ que forman una $CONS$. La proposición 1 nos permite concluir que, en efecto, cualquier $CONS$ provee una base de $L_2[0,1]$.

Proposición 1. Sea $\{f_j\}_{j=1}^{\infty}$ una $CONS$ de $L_2[0,1]$ y sea f una función de $L_2[0,1]$. Si definimos

$$\beta_j =< f, f_j >,\ j = 1, 2, \ldots \tag{2.53}$$

Entonces $\sum_{j=1}^{\lambda} \beta_j f_j$ es la mejor aproximación a f en el sentido que

$$\|f - \sum_{j=1}^{\lambda} \beta_j f_j\| \leq \|f - \sum_{j=1}^{\lambda} b_j f_j\|$$

para todo $\boldsymbol{b} = (b_1, b_2, \ldots, b_\lambda)^T \in \Re^\lambda$. Más aún, cuando $\lambda \to \infty$,

$$\|f - \sum_{j=1}^{\lambda} \beta_j f_j\| \to 0$$

Es decir, podemos representar f en $L_2[0,1]$ como una combinación lineal de las funciones de la base $\{f_j\}_{j=1}^{\infty}$. Llamaremos *coeficientes generalizados de Fourier* a los coeficientes β_j de la expresión 2.53. $\sum_{j=1}^{\infty} \beta_j f_j$ es la proyección de f sobre el subespacio generado por las funciones $\{f_j\}_{j=1}^{\infty}$. Llamaremos a $\sum_{j=1}^{\infty} \beta_j f_j$ la *expansión en series generalizadas de Fourier* de f.

Finalmente, se sabe que los coeficientes generalizados de Fourier satisfacen la relación de Parseval 2.54.

$$\sum_{j=1}^{\infty} \beta_j^2 = \|f\|^2 \tag{2.54}$$

2.5. Ejercicios

1. En el ejemplo 2.2.4 utilizamos la serie de cosenos para estimar la función de regresión. Otra posible CONS es la serie de senos en la que $f_1(x) = 1$ y $f_j(x) = \sqrt{2} sen(j\pi x))$, para $j = 2, 3, \ldots$. Ajuste el mejor modelo de series de senos a los datos del ejemplo 2.2.4.

2. Repita el ejercicio 1, pero usando la serie de senos y cosenos, en la que $f_1(x) = 1$, $f_{2j}(x) = \sqrt{2} cos(2j\pi x))$ y $f_{2j+1}(x) = \sqrt{2} sen(2j\pi x))$, para $j = 1, 2, \ldots$.

3. Ajuste un modelo de regresión polinómica a los datos de la Tabla 1.2 del ejercicio 2 del capítulo 1.

Capítulo 3

ESTIMADORES KERNEL

3.1. Introducción

Para establecer un marco general, asumiremos que disponemos de observaciones de la variable de respuesta Y para n valores predeterminados de una variable independiente X. Las n observaciones bivariadas disponibles, denotadas $(x_1, y_1), \ldots, (x_n, y_n)$, siguen el modelo

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{3.1}$$

donde $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ es un vector de errores aleatorios no correlacionados que tienen media cero y varianza común σ^2, f es una función de regresión desconocida y se satisface que $0 \leq x_1 < \cdots < x_n \leq 1$.

Para efectos de presentación de los resultados, asumiremos que los valores de X han sido elegidos así:

$$x_i = (2i - 1)/2n, \quad i = 1, 2, \ldots, n \tag{3.2}$$

Nuestro propósito es estimar f en 3.1, para lo cual buscaremos construir estimadores lineales que puedan escribirse en la siguiente forma general, que para un λ dado es una combinación lineal de las observaciones y_i, donde $K(\cdot, x_i; \lambda)$, $i = 1, \ldots, n$ es una colección de funciones ponderadoras que dependen de los x_i y de un parámetro de

suavización denotado λ:

$$f_\lambda(x) = \sum_{i=1}^{n} K(x, x_i; \lambda) y_i \tag{3.3}$$

3.2. Estimadores Kernel

En nuestra discusión sobre los estimadores de series encontramos que el estimador de cosenos (y se aplica en general para todos los estimadores de series) puede representarse como un estimador lineal, que se construye con una colección de funciones de pesos, una para cada x_i, que depende de λ. La función de asignación de pesos 2.45 que utilizamos en este caso tiene la limitación de oscilar a medida que se aleja del punto de estimación x, de tal manera que asigna pesos bajos a algunos puntos cercanos al punto de estimación y pesos más altos a puntos más alejados de este punto de estimación (ver figura 2.8).

Una opción de construcción de estimadores que eviten esta limitación sería utilizar la misma idea pero con una función ponderadora que asigne consistentemente pesos más bajos a observaciones más alejadas del punto de estimación, al mismo tiempo que asigna pesos altos a las observaciones cercanas al punto de estimación.

Una manera de producir estos estimadores es utilizar colecciones de funciones ponderadoras que tengan la forma general:

$$K(x, x_i; \lambda) = \frac{1}{\lambda} K\left(\frac{x - x_i}{\lambda}\right), i = 1, \ldots, n \tag{3.4}$$

con K una función simétrica con soporte en $[-1, 1]$ que tiene su máximo en cero. Al parámetro λ, que en los estimadores de series era el número de términos de la serie que se conservaban en la estimación de f, lo llamaremos indistintamente *parámetro de suavización* o *ancho de banda* y cumplirá la misma función: servir de base para elegir *el mejor* estimador de f con los criterios de bondad discutidos en la sección 2.2.3. Pero a diferencia de los estimadores de series, en los estimadores kernel λ no tiene que ser un entero, aunque sí restringiremos su valor a que sea cualquier real no negativo.

Con este modo de hacer, la observación y_i asociada con un x_i cercano a x tendrá asignado un peso más alto, contribuyendo de esta manera lo máximo posible a la estimación de la función de regresión

en x. Al mismo tiempo, observaciones y_i asociadas con x_i alejadas de x tendrán pesos cada vez más pequeños a medida que estén más alejados de x.

Para garantizar lo anterior, le exigiremos por lo pronto a la función K que tenga las siguientes dos propiedades:

$$\int_{-1}^{1} K(u)du = 1$$
$$\int_{-1}^{1} uK(u)du = 0 \tag{3.5}$$

La primera propiedad en 3.5 ayuda a garantizar que la suma de los pesos sea igual a 1; la segunda, a que K sea simétrica alrededor de cero.

A estas funciones las llamaremos *funciones kernel*. Y a los estimadores basados en estas funciones los llamaremos *estimadores kernel*.

Dos funciones Kernel de uso común, que se conocen como kernel cuadrático (*Epanechnikov*) y kernel biponderado (*biweight*) se describen en la tabla 3.1 y se ilustran en la figura 3.1.

Tabla 3.1: Dos kernel de uso común en la construcción de estimadores kernel

Kernel	K
Cuadrático	$K(u) = \frac{3}{4}(1-u^2)I_{[-1,1]}(u)$
Biponderado	$K(u) = \frac{15}{16}((1-u^2)^2)I_{[-1,1]}(u)$

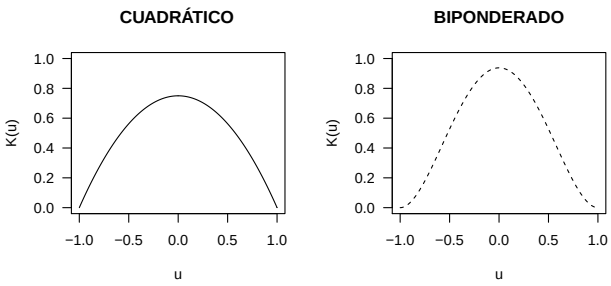


Figura 3.1: Kernel cuadrático y kernel biponderado

3.2.1. Estimador kernel de Priestley-Chao

Con el estimador de cosenos concluimos que una forma particular del estimador es:

$$f_\lambda(x) = \frac{1}{n}\sum_{i=1}^{n} K(x, x_i; \lambda) y_i \tag{3.6}$$

Si incorporamos ahora nuestro kernel genérico 3.4 obtendremos el estimador kernel

$$f_\lambda(x) = \frac{1}{n\lambda}\sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right) y_i \tag{3.7}$$

El estimador kernel 3.7 es un estimador kernel genérico que luce tal como el estimador de cosenos 2.44 pero cuya función de asignación de pesos luce muy diferente de la función ponderadora 2.45 ya que si bien coinciden en asignar pesos altos en puntos cercanos al punto de estimación, asigna cada vez pesos menores a observaciones alejadas de x.

Este estimador fue propuesto por Priestley & Chao (1972) para el caso de diseños igualmente espaciados como el del diseño 2.2. Sin embargo su desempeño no es muy bueno, como se ve en la figura 3.2, en la que se destaca la escasa habilidad del estimador en los extremos del diseño.

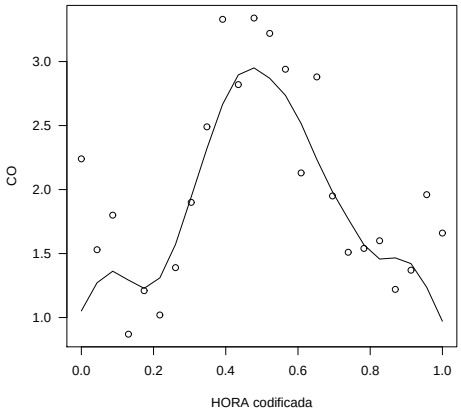


Figura 3.2: Estimación kernel del comportamiento diario del CO horario en la calle 15 de Cali usando el Kernel genérico con el kernel cuadrático y con $\lambda = 0.15$

El ancho de banda se elige de manera similar, basado en la minimización del criterio de validación cruzada generalizada GCV.

En este caso el estimador lucirá como $\boldsymbol{f_\lambda} = \boldsymbol{S_\lambda y}$, donde un término S_{ij} de la matriz $\boldsymbol{S_\lambda}$ lucirá como:

$$S_{ij} = (n\lambda)^{-1} K(\lambda^{-1}(x_j - x_i)), \quad i, j = 1, \ldots, n$$

Otro aspecto importante es la selección del kernel. Una alternativa de selección sería identificar el kernel que minimiza el riesgo de f_λ, una vez elegido el ancho de banda. Estudios previos (Benedetti 1975, Gasser & Müller 1979) identifican el kernel cuadrático, propuesto por Epanechnikov (1969) y adaptado por Gasser & Müller (1979) para que sea positivo en $[-1, 1]$, como la mejor elección entre varios kernel sometidos a prueba por estos autores. Nuevos kernel han sido propuestos posteriormente, tales como el kernel Gaussiano

$$K(u) = (2\pi)^{-1/2} e^{-u^2}, u \in (-\infty, \infty)$$

que algunos autores (Diggle, Heagerty, Liang & Zeger 2002) limitan a la expresión más simple

$$K(u) = e^{-u^2}, u \in (-\infty, \infty)$$

y el kernel triponderado (triweight)

$$35/32(1 - u^2)^3, u \in [-1, 1],$$

pero en general la elección del kernel tiene baja influencia, comparada con la elección del ancho de banda, en la calidad de los estimadores.

3.2.2. Estimador de Nadaraya-Watson

Varios intentos posteriores tratan de mejorar la habilidad del estimador genérico. El primero fue propuesto por Benedetti (1975), en el marco en el que se desarrolla este texto, con un diseño de puntos tal como el diseño 2.2. El mismo fue propuesto por Nadaraya & Seckler (1964) y Watson (1964), para el caso en el que X es una variable aleatoria. Este estimador de conoce como *estimador de Nadaraya-Watson* y se define así:

$$f_\lambda(x) = \frac{\sum_{i=1}^{n} K(\lambda^{-1}(x - x_i)) y_i}{\sum_{j=1}^{n} K(\lambda^{-1}(x - x_j))} \tag{3.8}$$

El estimador de Nadaraya-Watson ajusta los pesos asignados por el kernel de tal manera que sumen uno, haciendo que el estimador

sea realmente un promedio ponderado de las observaciones cercanas a cada punto de estimación.

En el lenguaje **R** la estimación kernel es parte de la distribución base en el paquete *STATS* (R Development Core Team 2011) y se ejecuta con la función *ksmooth*, que utiliza el estimador de Nadaraya-Watson. La figura 3.3 muestra la estimación kernel del comportamiento diario del CO horario en la calle 15 de Cali usando la función *ksmooth* de **R**, con $\lambda = 0.15$. Esta estimación contrasta con la estimación representada en la figura 3.2, en la que el ajuste se hace con el estimador de Priestley-Chao, particularmente en los extremos del diseño en los que el estimador de Nadaraya-Watson se desempeña mucho mejor.

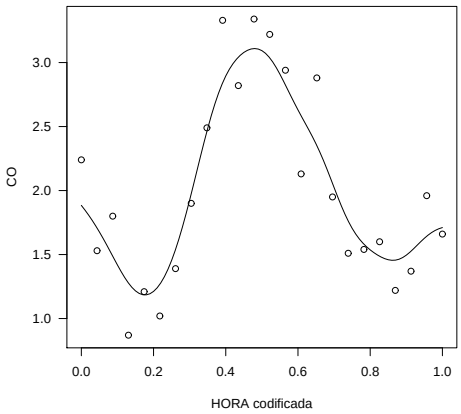


Figura 3.3: Estimación kernel del comportamiento diario del CO horario en la calle 15 de Cali usando la función *ksmooth* de **R**, con $\lambda = 0.15$. La función *ksmooth* usa el estimador de Nadaraya-Watson

La versión original del estimador de Nadaraya-Watson, que presume que X es una variable aleatoria, tiene una sólida justificación apoyada en el modelo (1.11), en el que tenemos que

$$f(x) = E(Y|X = x)$$

En este caso requeriríamos la distribución condicional de Y dado X, digamos $h(y|x)$, de tal manera que

$$f(x) = \int yh(y|x)dy = \int y\frac{h(x,y)}{h_X(x)}dy \tag{3.9}$$

donde $h(x, y)$ es la distribución conjunta de X y Y, con $h_X(x)$ la distribución marginal de X.

Un solución para este problema sería utilizar estimación de densidades vía suavización. Este es un tema presente en la literatura por muchos años. Härdle (1990*b*) y Simonoff (1996) incluyen una extensa discusión en los capítulos iniciales de sus libros, dedicados al estudio de los problemas de suavización, a la estimación de densidades. Härdle, por ejemplo, incluye una extensa discusión sobre estimación kernel de densidades, sobre estimación de densidades basada en series ortogonales y sobre estimadores de densidades por máxima verosimilitud penalizada. Simonoff por su parte sigue aproximadamente el mismo rumbo y cubre el problema de la estimación de densidades en sus cuatro primeros capítulos.

Un estimador kernel-producto de $h(x, y)$ es

$$\hat{h}(x,y) = \frac{1}{\lambda_x \lambda_y} \sum_{i=1}^{n} K_x(\frac{x - x_i}{\lambda_x}) K_y(\frac{y - y_i}{\lambda_y}) \tag{3.10}$$

Y un estimador de $h_X(x)$ es

$$\hat{h}(x) = \frac{1}{\lambda_x} \sum_{i=1}^{n} K_x(\frac{x - x_i}{\lambda_x}) \tag{3.11}$$

Si reemplazamos los estimadores 3.10 y 3.11 en 3.9 y consideramos las propiedades 3.5, se reconstruye el estimador de Nadaraya-Watson 3.8 (Ejercicio 1).

3.2.3. Estimador de Gasser-Müller

Otro estimador kernel común fue propuesto por Gasser & Müller (1979), el cual tiene la siguiente forma:

$$f_\lambda(x) = \sum_{i=1}^{n} \left(\lambda^{-1} \int_{s_{i-1}}^{s_i} K(\lambda^{-1}(x - s)) ds \right) y_i \tag{3.12}$$

con

$$s_0 = 0, s_{i-1} \leq x_i \leq s_i, i = 1, \ldots, n-1, s_n = 1$$

Tenemos entonces tres estimadores kernel: el estimador de Priestley-Chao 3.7, el estimador de Nadaraya-Watson 3.8 y el estimador de Gasser-Muller 3.12. Estos estimadores tienen aproximadamente las mismas propiedades asintóticas, por lo que en lo sucesivo no nos detendremos a identificar cuál de ellos está en

uso. Existe una versión del estimador de Piestley-Chao para diseños desigualmente espaciados, mientras que el estimador de Gasser-Muller está diseñado específicamente para este tipo de diseños desigualmente espaciados. El estimador de Nadaraya-Watson fue pensado por sus autores para diseños aleatorios y ha tomado la delantera entre la mayoría de los usuarios de los métodos de regresión no paramétrica por su flexibilidad y eficiencia.

3.2.4. Estimadores lineales localmente

Otro estimador común, que también es un estimador kernel, es el *estimador de regresión local* (Cleveland 1979). Se sabe (Eubank (1999, pág. 189), Simonoff (1996, pág. 138)) que el estimador de Nadaraya-Watson 3.8 coincide con la solución para β_0 del problema de mínimos cuadrados ponderados

$$\sum_{i=1}^{n}(y_i - \beta_0)^2 K\left(\frac{x - x_i}{\lambda}\right), \tag{3.13}$$

lo que sugiere el ajuste de polinomios locales de orden p, por ejemplo uno que minimice

$$\sum_{i=1}^{n}(y_i - \beta_0 - \ldots - \beta_p(x - x_i)^p)^2 K\left(\frac{x - x_i}{\lambda}\right) \tag{3.14}$$

Al estimador resultante de minimizar la suma 3.14 se le conoce como *estimador polinomial de regresión local de orden p.*

Ilustraremos el caso de la solución del sistema 3.13, el cual se ha popularizado con el nombre de *estimador LOESS*, que utiliza los k vecinos más cercanos y estima el valor de la función de regresión en el punto x conforme a los siguientes pasos:

1 Identificar los k vecinos más cercanos de x y denotar este conjunto como $N(x)$

2 Encontrar $\Delta(x) = max_{\{x_i \in N(x)\}}|x - x_i|$ (la distancia asociada con el vecino mas cercano que se encuentra más alejado de x).

3 Asignar pesos w_i a cada punto en $N(x)$ usando la función de pesos tri-cubo

$$W\left(\frac{|x - x_i|}{\Delta(x)}\right)$$

donde

$$W(u) = \begin{cases} (1-u^3)^3, & \text{para } 0 \leq u \leq 1 \\ 0, & \text{en cualquier otro caso} \end{cases}$$

4 Ajustar una recta por mínimos cuadrados ponderados de Y sobre X, confinada al conjunto $N(x)$, utilizando los pesos obtenidos en 3.

La estimación lineal local LOESS de $f_\lambda(x)$ en el punto x toma como valor el del término independiente de la la recta de regresión local resultante. Es decir, si la recta ajustada es $\beta_{\lambda 0} + \beta_{\lambda 1} x$, entonces $f_\lambda(x) = \beta_{\lambda 0}$.

En la figura 3.4 aparece la curva estimada usando LOESS. En este caso el grado de suavización se controla usando la proporción de observaciones más cercanas. Para construir la estimación loess ilustrada en la figura 3.4 se utilizó la proporción 0.5.

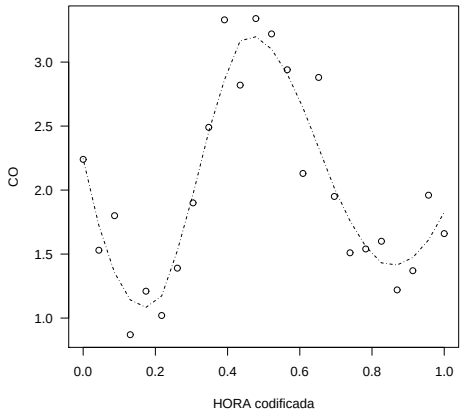


Figura 3.4: Estimación lineal local del comportamiento diario del CO horario en la calle 15 de Cali usando la función *loess* de **R**. La función *loess* usa el kernel Gausiano

3.2.5. Estimación kernel multivariante

Todos los estimadores kernel que hemos discutido en esa sección pueden ampliarse a la estimación de funciones p-variadas. Supongamos que $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^T$. La solución es definir los kernel multivariados como productos de kernel univariados, así:

$$K(u_1, \ldots, u_p) = \prod_{j=1}^{p} K_j(u_j) \tag{3.15}$$

De esta manera, el estimador de Priestley-Chao, versión multivariada, tomaría la siguiente forma:

$$f_\lambda(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} y_i \prod_{h=1}^{p} \lambda_h^{-1} K_h(\lambda_h^{-1}(x_h - x_{ih})) \tag{3.16}$$

Como se puede anticipar, la estimación multivariante tiene sus propias dificultades. Por ejemplo, en el caso de la observación de la contaminación horaria de CO en un día típico en Cali tenemos 24 mediciones de CO, una para cada hora del día, lo que implica un diseño con 24 intervalos del mismo tamaño. Supongamos ahora que disponemos de una segunda variable de predicción, digamos la temperatura promedio horaria. Si construimos un diseño de 10 subintervalos de la misma longitud para la variable temperatura, tendríamos entonces 240 subintervalos para tratar de construir un modelo bivariado con 24 respuestas.

Así que es posible generalizar la estimación no-paramétrica al caso p-dimensional. Sin embargo, como lo ilustra el ejemplo del CO en Cali, es necesario enfatizar tres dificultades que surgen en en el problema de la estimación no-paramétrica multivariante:

1 Los estimadores multivariantes son necesariamente más complicados que los univariados. En la práctica hay muchas más opciones que elegir y se deben escoger más parámetros de suavización.

2 La visualización gráfica es difícil en la estimación multivariante. Para el caso de una variable explicativa bidimensional es aún posible obtener alguna visualización. Pero ciertamente no lo es para variables explicativas en tres o más dimensiones.

3 A medida que la dimensión de la variable explicativa se incrementa, los estimadores son cada vez más inexactos. Supóngase que se alcanza una exactitud dada en la estimación de una función de regresión en un espacio k-dimensional. Entonces, en dimensiones mayores que k, se necesitan muestras mucho más grandes para alcanzar la misma exactitud. Este hecho se conoce como *la maldición de la dimensionalidad*. Una consecuencia de ello es que, en dimensiones altas, vecindades "locales" están en casi todas vacías y vecindades que no están vacías, no son en general "locales".

Considérese el siguiente ejemplo (Simonoff 1996). Supóngase que se tiene una muestra uniforme en el hipercubo $[-1, 1]$. 79 % de las observaciones caerán en el círculo unitario centrado en el origen cuando $p = 2$, pero solamente 15 % si $p = 5$ y 0.25 % si $p = 10$. Es decir, vecindades grandes prácticamente no incluyen observaciones, implicando la pérdida del caracter local de la estimación.

La forma general de un estimador kernel multidimensional es

$$f_L(\mathbf{x}) = \frac{1}{n|\mathbf{L}|} \sum_{i=1}^{n} K_p[\mathbf{L}^{-1}(\mathbf{x} - \mathbf{x}_i)] \tag{3.17}$$

donde $K_p : \Re^p \to \Re$ es la función kernel (la cual es en general una función de densidad; $\mathbf{L}$ es una matriz $p \times p$ de anchos de banda, no-singular; y $|\mathbf{L}|$ es el valor absoluto del determinante de la matriz $\mathbf{L}$. La técnica más común para obtener K_p a partir de una función kernel univariada K es utilizar un producto de la forma 3.15.

Un ejemplo de estimación no paramétrica bivariada se presenta en la figura 3.5, en la que se estima el comportamiento del contaminate Ozono (O_3) dependiendo de la temperatura y la humedad relativa. Las mediciones son promedios diarios y corresponden a la ciudad de Rio de Janeiro entre el primero de enero de 2001 y el 31 de diciembre de 2005 (Junger & de Leon 2011).

En este problema de la contaminación por ozono troposférico en Rio de Janeiro entre 2001 y 2005, hemos ajustado un modelo bivariado de la forma:

$$Y = f(X_1, X_2) + \epsilon, \tag{3.18}$$

donde Y representa el promedio diario de Ozono y X_1 y X_2 los promedios diarios de temperatura y humedad, respectivamente.

Los resultados indican que las concentraciones diarias más altas de Ozono se produjeron durante días en ese periodo con altas temperaturas asociadas con altas humedades relativas. La respuesta luce muy lineal, por lo que, sujeto al cumplimiento de otros supuestos, un modelo lineal podría ser adecuado para el análisis conjunto de estas variables. La función *sm.regression* de **R** se debe a Bowman & Azzalini (2010).

3.3. Inferencia en la estimación kernel

Nuestro propósito es construir inferencias sobre la función de regresión f a partir de nuestro estimador kernel f_λ. Ya sabemos que

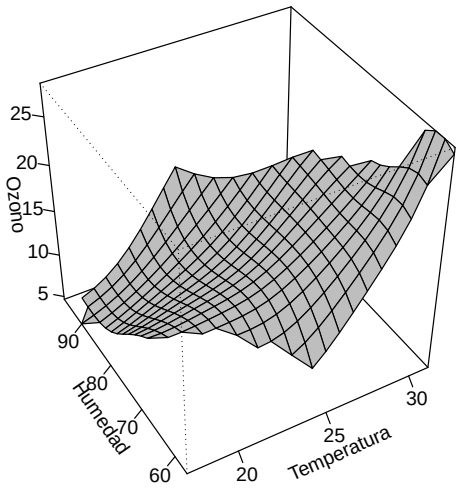


Figura 3.5: Estimación bivariada del comportamiento del Ozono troposférico (promedio diario), dependiendo de la temperatura y la humedad promedio diarias en Rio de Janeiro en el periodo 2001-2005, usando la función *sm.regression* de **R**.

f_λ es un estimador sesgado, por lo que intuimos que un intervalo de confianza para f construido a partir de f_λ deberá considerar no solamente la varianza del estimador, si no también su sesgo. Previamente, sin embargo, deberemos asegurarnos de que f_λ es en realidad un estimador consistente de f. Así que discutiremos primero el problema de la consistencia, para avanzar hacia la estimación por intervalos.

3.3.1. Consistencia de los estimadores kernel

Como f_λ es sesgado, su consistencia conviene estudiarla a partir del riesgo, tal como anticipamos en la sección 2.2.3. Es decir, nos gustaría hallar una expresión del comportamiento asintótico de

$$R(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E[f(x_i) - f_\lambda(x_i)]^2.$$

Una forma de lograrlo es trabajar primero sobre la forma asintótica del riesgo en un punto x en $[0, 1]$ y luego tratar de obtener la suma para todo x en el diseño de puntos.

La expresión para el riesgo en un punto x sería:

$$R(x; \lambda) = E[f(x) - f_\lambda(x)]^2 = \{f(x) - E[f_\lambda(x)]\}^2 + \mathrm{var}[f_\lambda(x)] \quad (3.19)$$

por lo que podríamos estudiar por aparte el comportamiento el sesgo y de la varianza y luego sumar para obtener el riesgo.

Ilustraremos aquí la consistencia para el estimador de Gasser-Müller, basados en las propuestas de Eubank (1999). En Härdle (1990*a*) se ilustra la consistencia para el estimador de Nadaraya-Watson.

Sabemos que las funciones K tienen soporte en $[-1, 1]$ y que su máximo en cero. Para continuar, necesitaremos las siguientes dos nuevas propiedades de las funciones kernel, adicionales a las requeridas en 3.5.

$$\begin{aligned} M_2 &= \int_{-1}^{1} u^2 K(u) du \neq 0 \\ V &= \int_{-1}^{1} K(u)^2 du < \infty \end{aligned} \tag{3.20}$$

Sabemos que para el estimador de Gasser-Müller la varianza de f_λ es:

$$\operatorname{var}(f_\lambda) = \frac{\sigma^2}{\lambda^2} \sum_{i=1}^{n} \left[\int_{s_{i-1}}^{s_i} K\left(\frac{1}{\lambda}(x-u)\right) du \right]^2$$

que puede expresarse como

$$\begin{aligned} \operatorname{var}(f_\lambda) &= \frac{\sigma^2}{n\lambda^2} \int_0^1 K^2\left(\frac{1}{\lambda}(x-u)\right) du + O((n\lambda)^{-2}) \\ &= \frac{\sigma^2 V}{n\lambda} + O((n\lambda)^{-2}) \end{aligned} \tag{3.21}$$

donde V se definió en 3.20 (ver Eubank (1999, Pág. 166)).

Por otra parte, si $f \in C^2[0, 1]$, introducimos un cambio de variable, usamos una expansión de Taylor y asumimos que f'' es continua, entonces:

$$\begin{aligned} E[f_\lambda(x)] &= \frac{1}{\lambda} \int_0^1 K(\lambda^{-1}(x-s)) ds + O(n^{-1}) \\ &= \int_{(x-1)/\lambda}^{x/\lambda} K(u) f(x-u) du, \quad j = 0, 1, 2 \\ &= f(x) M_0(x) - \lambda f'(x) M_1(x) + (\lambda^2/2) f''(x) M_2(x) + o(\lambda^2) \end{aligned} \tag{3.22}$$

donde

$$M_j(x) = \int_{(x-1)/\lambda}^{x/\lambda} u^j K(u)du$$

Si λ es suficientemente pequeño, entonces $[(x-1)/\lambda, x/\lambda] \supset [-1, 1]$ por lo que a partir de las condiciones 3.5 tenemos que:

$$\begin{aligned} M_0 &= 1 \\ M_1 &= 0 \\ M_2 &= M_2 \end{aligned}$$

Se sigue que el sesgo de f_λ sería (Eubank 1999, Pág. 167):

$$\begin{aligned} B[f_\lambda(x)] &= E[f_\lambda(x) - f(x))] \\ &= (\lambda^2/2) f''(x) M_2 + o(\lambda^2) + O(n^{-1}) \end{aligned} \tag{3.23}$$

Tenemos que aquí que:

1. La primera condición en 3.5 garantiza que el primer término de $E[f_\lambda]$ en 3.22 es $f(x)$.

2. La segunda condición en 3.5 permite suprimir el término de orden λ en el sesgo (Ver Ejercicio 2 propuesto al final del capítulo), asegurando que es $O(\lambda^2 + n^{-1})$.

3. Las implicaciones de la primera condición en 3.20 son más complejas, ya que esta elección hace que K se asocie con una función en $C^2[0, 1]$ en el sentido que su sesgo se hace decrecer tan rápidamente como es posible para una función de regresión que se sabe tiene dos derivadas continuas.

4. La segunda condición en 3.20 se requiere para garantizar que f_λ tenga varianza finita.

Usando el sesgo 3.23 al cuadrado y la varianza 3.21 de f_λ podemos regresar a nuestra evaluación del riesgo en un punto x 3.19, que sería:

$$R(\lambda) = \frac{\sigma^2 V}{n\lambda} + \frac{\lambda^4}{4} f''(x)^2 M_2^2$$

Pero en realidad necesitamos extender el riesgo a todos los puntos en el diseño, lo que luce sencillo si uno intenta obtener el riesgo sumando sobre todos los x_i. Sin embargo, tropezamos con una dificultad debida

a que en los extremos del diseño (es decir en $x \in [0, \lambda]$ y en $x \in [1-\lambda, \lambda]$) no se puede garantizar que $M_0(x) = 1$, lo que sí se puede garantizar en el interior del diseño (ver Ejercicio 3 propuesto al final del capítulo). En tal caso, una opción sería transformar la función kernel con el propósito de mejorar la estimación en estos bordes.

Supongamos que nos situamos en el extremo inferior $x \in [0, \lambda]$ y definamos $0 \leq q < 1$. Como los valores de x en este intervalo son fracciones de λ, siempre será posible escribir $x = q\lambda$. En estos extremos la función kernel cuadrático de la Tabla 3.1 podría reescribirse como:

$$K_q(u) = \frac{12(u+1)}{(1+q)^4}\{(1-2q)u + 0.5(3q^2 - 2q + 1)\}I_{[-1,q]}(u) \quad (3.24)$$

de tal manera que se satisfacen todas las propiedades 3.5 y 3.20 de las funciones kernel.

La figura 3.6 ilustra cómo luce la función kernel cuadrático (caso $q = 1$) de la Tabla 3.1 junto con dos versiones de la función kernel cuadrático de borde del tipo 3.24 para $q = 1/5$ y $q = 2/5$.

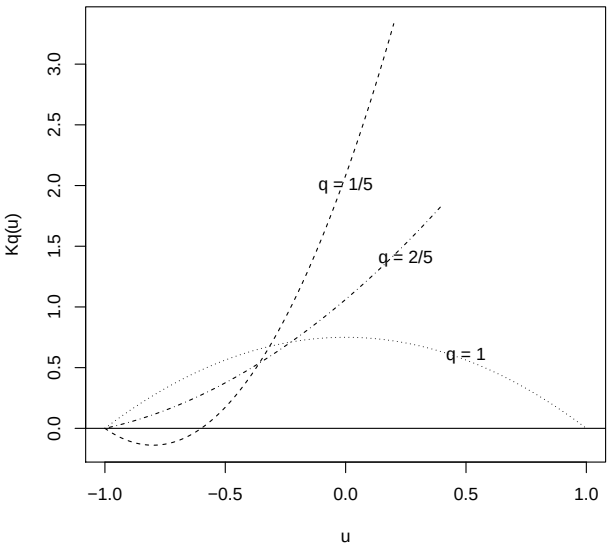


Figura 3.6: Función kernel cuadrático (caso $q = 1$) junto con dos versiones de la función kernel cuadrático de borde del tipo 3.24 para $q = 1/5$ y $q = 2/5$

Utilizando funciones kernel de borde con el estimador de Gasser-Müller se llega que el riesgo a lo largo del diseño será:

$$R(\lambda) = \frac{\sigma^2 V}{n\lambda} + \frac{\lambda^4}{4} J_2(f) M_2^2$$

con

$$J_2(f) = \int_0^1 f''(x)^2 dx$$

de tal manera que la tasa de convergencia del estimador, si se usan funciones kernel de borde en los extremos del diseño, podría alcanzar el óptimo $n^{4/5}$, cuando se usa un estimador lineal y se asume que $f \in W_2^2[0,1]$.

3.3.1.1. Dos anotaciones

Un comentario necesario en este momento es que los estimadores de series y los estimadores kernel son métodos diferentes de estimación, a pesar de que la motivación para proponer los estimadores kernel surja de una representación particular de la serie de cosenos usando un kernel Dirichlet.

Y por otra parte, hemos discutido estos resultados en el marco de un diseño uniforme, pero no hay nada que diga que no son aplicables para diseños no uniformes. De hecho la discusión principal usa el estimador de Gasser-Müller, que no requiere de un diseño uniforme.

3.3.2. Estimación por intervalos

Los estimadores kernel pueden verse como promedios ponderados, por lo que el Teorema del Límite Central es aplicable y la distribución de f_λ será asintóticamente normal. Un posible enunciado para este resultado sería:

Proposición 2. Si los ϵ_i son independientes e idénticamente distribuidos con $E(\epsilon_i) = 0$ y $\text{var}(\epsilon_i) = \sigma^2 < \infty$ y si $n \to \infty$ y $\lambda \to 0$ de tal modo que $n\lambda \to \infty$, entonces para $x \in [0,1]$ se satisface que

$$\frac{f_\lambda(x) - E[f_\lambda(x)]}{\sqrt{\text{var}[f_\lambda(x)]}} \xrightarrow{d} N[0,1]$$

donde $N[0,1]$ representa una variable aleatoria normal estándar.

En el enunciado 2 se acude únicamente al hecho de que f_λ es un promedio y el resultado no depende de suavización alguna, porque $f_\lambda(x) - E[f_\lambda(x)]$ no depende de f. El problema es más complejo cuando nos enfrentamos al estudio de la normalidad asintótica de la cantidad $f_\lambda(x) - f(x)$, que sí requiere del uso de suavización. En este caso

acudiremos al resultado:

$$f_\lambda(x) - f(x) = \{f_\lambda(x) - E[f_\lambda(x)]\} + \{E[f_\lambda(x)] - f(x)\} \qquad (3.25)$$

Si dividimos ambos lados de la expresión 3.25 entre $\text{var}[f_\lambda(x)]$, entonces el problema de convergencia se reduce al segundo término de la derecha, que tiende a cero asintóticamente si $n\lambda^5 \to 0$, por lo que podemos expresar el siguiente nuevo enunciado:

Proposición 3. Bajo las mismas condiciones del enunciado 2, con $f \in C^2[0,1]$ y $n\lambda^5 \to 0$, entonces:

$$\frac{f_\lambda(x) - f(x)}{\sqrt{\text{var}[f_\lambda(x)]}} \xrightarrow{d} N[0,1]$$

Llegamos entonces a un primer intervalo de confianza para $f(x)$:

$$f_\lambda(x) \pm Z_{\alpha/2}\hat{\sigma}\sqrt{\sum_{i=1}^{n} w_{in}^2(x;\lambda)} \qquad (3.26)$$

Para el caso del estimador de Gasser-Müller tenemos que:

$$w_{in}(x;\lambda) = \lambda^{-1}\int_{s_{i-1}}^{s_i} K(\lambda^{-1}(x.s))ds, \quad i = 1,\ldots,n$$

Y para el caso del estimador de Nadaraya-Watson,

$$w_{in}(x;\lambda) = \frac{K(\lambda^{-1}(x - x_i))}{\sum_{j=1}^{n} K(\lambda^{-1}(x - x_j))}, \quad i = 1,\ldots,n$$

En el intervalo 3.26, $Z_{\alpha/2}$ es el percentil $100(1-\alpha)$ de la normal estándar y $\hat{\sigma}$ es la raíz cuadrada de una estimación $\hat{\sigma}^2$ de σ^2 basada en alguno de los estimadores de la varianza en el modelo 2.1 presentados en la sección 2.2.1.

La principal limitación del intervalo 3.26 es que se construye a partir de un estimador f_λ de f que se sabe que no es insesgado. Vale recordar que hemos elegido λ del tal manera que se minimice el riesgo, que es una combinación del sesgo y de la varianza. Una opción sería intentar construir un intervalo de confianza que considere el sesgo.

Eubank (1999) propone una solución que implica estimar f con una parámetro de suavización, digamos λ_1, estimar luego f'' usando

otro parámetro de suavización, digamos λ_2, y usar funciones kernel de borde en los extremos. El intervalo tomaría la siguiente forma:

$$f_{\lambda_1}(x) - \frac{\lambda_2}{2} M_2 f''_{\lambda_2}(x) \pm Z_{\alpha/2}\hat{\sigma}\sqrt{\sum_{i=1}^{n} w_{in}^2(x;\lambda)} \tag{3.27}$$

En esta propuesta se debe además adecuar el cálculo de los pesos $w_{in}(x;\lambda)$ para considerar la forma estimación de f y tiene como dificultades la elección de más de un parámetro de suavización y el uso de funciones kernel particulares en los extremos. Por otra parte, el intervalo puede verse como si se centrara en una corrección del estimador para lograr que sea insesgado.

Bowman & Azzalini (1997) prefiere utilizar "bandas de variabilidad" que se construyen usando intervalos de confianza calculados para valores de x, alrededor de $E[f_\lambda(x)]$ usando un estimador de σ^2.

La figura 3.7 muestra los intervalos de confianza construidos usando un estimador kernel y las bandas de confianza de Bowman-Azzalini. Ambos ajustes tienen aproximadamente el mismo número de grados de liberta equivalente.

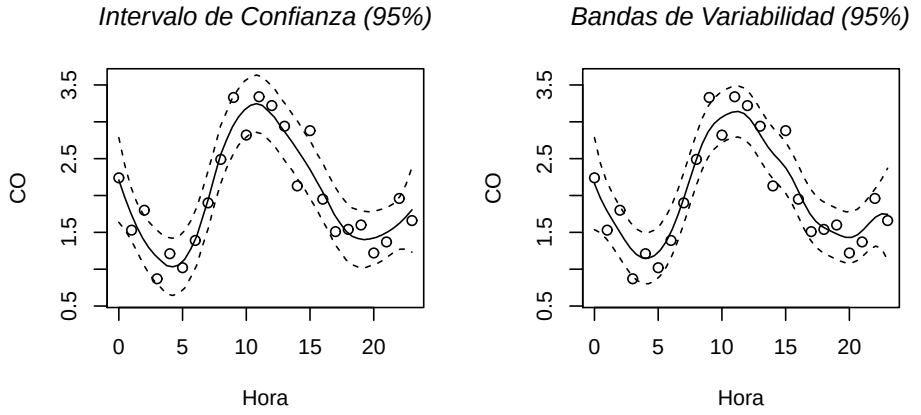


Figura 3.7: Intervalos (punto a punto) del 95 % de confianza para $f(x)$ y bandas de variabilidad para $E[f(x)]$, para los datos de CO del ejemplo 2.2.4

3.4. Ejercicios

1. Verifique que reemplazando los estimadores 3.10 y 3.11 en 3.9 y considerando las propiedades 3.5, se reconstruye el estimador de Nadaraya-Watson.

2. Pruebe que la segunda condición en 3.5 permite suprimir el término de orden λ en el sesgo, asegurando que es $O(\lambda^2 + n^{-1})$.

3. Sustente las razones por las cuales si bien se puede garantizar en el interior del diseño que $M_0(x) = 1$, en los extremos del diseño (es decir en $x \in [0, \lambda]$ y en $x \in [1 - \lambda, \lambda]$) esto no se puede garantizar.

PÁGINA EN BLANCO
EN LA EDICIÓN IMPRESA

Capítulo 4

ESTIMACIÓN SPLINE

4.1. Introducción

Para establecer un marco general, asumiremos que disponemos de observaciones de la variable de respuesta Y para n valores predeterminados de una variable independiente X. Las n observaciones bivariadas disponibles, denotadas $(x_1, y_1), \ldots, (x_n, y_n)$, siguen el modelo

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{4.1}$$

donde $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ es un vector de errores aleatorios no correlacionados que tienen media cero y varianza común σ^2, f es una función de regresión desconocida y se satisface que $0 \leq x_1 < \cdots < x_n \leq 1$.

Para efectos de presentación de los resultados, asumiremos que los valores de X han sido elegidos así:

$$x_i = (2i - 1)/2n, \quad i = 1, 2, \ldots, n \tag{4.2}$$

Nuestro propósito es estimar f en 4.1, para lo cual buscaremos construir estimadores lineales que puedan escribirse en la siguiente forma general, que para un λ dado es una combinación lineal de las observaciones y_i, donde $K(\cdot, x_i; \lambda)$, $i = 1, \ldots, n$ es una colección de funciones ponderadoras que dependen de los x_i y de un parámetro de

suavización denotado λ:

$$f_\lambda(x) = \sum_{i=1}^{n} K(x, x_i; \lambda) y_i \tag{4.3}$$

4.2. Interpolación y suavización spline

Aunque esta sección tiene como objetivo el estudio de la suavización spline, nos ayudará mucho en esta parte una rápida revisión del tema (estrechamente ligado) de interpolación.

Mientras que la interpolación es un tema de estudio del análisis numérico, la suavización spline es un tema estadístico. Lo que haremos será tomar prestadas algunas ideas tomadas de la interpolación en nuestro camino hacia el estudio de la suavización spline.

En la interpolación, si se sabe que tenemos n observaciones $(x_1, y_1), \ldots (x_n, y_n)$ y que todas provienen de un polinomio, es posible reconstruir este polinomio para calcular los valores de Y para valores "no observados" de X. Las técnicas de reconstrucción de polinomios no son un propósito principal de este texto, pero se discutirán dos de ellas en las secciones 4.2.1 y 4.2.2, porque permitirán identificar algunos elementos teóricos útiles en la estimación de funciones de regresión propios del análisis estadístico, en el que el interés se centra más en la estimación que en la interpolación, por cuanto se asume que las observaciones se ajustan al modelo 4.1.

4.2.1. Interpolación

Si se dispone de un conjunto A de k puntos en $\Re^2$, es bien sabido que podrá ajustar a ellos exactamente un polinomio de grado $k-1$. Es decir, con dos puntos se puede ajustar una recta y con cuatro puntos una cúbica. Sea $A = (p_1, p_2, \ldots, p_k)$ el conjunto de k puntos del polinomio π_{k-1} que se desea interpolar. Y sean $(x_i, y_i), i = 1, 2, \ldots, k$ las coordenadas de los puntos del conjunto A. El polinomio π_{k-1} se puede representar como se describe en la Ecuación 4.4.

$$\pi_{k-1} = a_0 + a_1 x + a_1 x^2 + \ldots + a_{k-1} x^{k-1} \tag{4.4}$$

Para determinar completamente el polinomio π_{k-1} basta con encontrar los coeficientes a_0, a_1, a_1,$\ldots$,a_{k-1}, para lo cual se utilizan las coordenadas $(x_i, y_i), i = 1, 2, \ldots, k$ de los k puntos de A. Se requieren

k ecuaciones, una para cada coeficiente, las que se construyen usando los k puntos de A.

Supongamos, por ejemplo, que tenemos

$$A = \{(1, 2); (2, 5); (3, 4); (4, 7)\}$$

y sea $\pi_3 = y = f(x) = a_0 + a_1x + a_2x^2 + a_3x^3$ el polinomio cuya gráfica pasa por estos cuatro puntos. Para identificar el polinomio π_3 usando interpolación, basta con encontrar los valores de las constantes a_0, a_1, a_2 y a_3. Construimos y resolvemos el sistema de ecuaciones lineales y encontramos la solución $\gamma^T = (-13, 23\frac{2}{3}, -10, 1\frac{1}{3})$. El gráfico del polinomio aparece en la Figura 4.1.

El procedimiento descrito para recuperar el polinomio π_3 se llama *interpolación* y presume que los cuatro puntos del conjunto A están sobre el polinomio de grado 3.

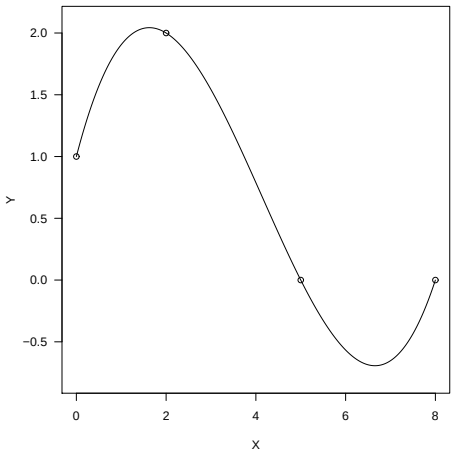


Figura 4.1: Gráfico del polinomio $y = -13 + 23 \cdot x - 10x + 1 \cdot x$

4.2.2. Interpolación por partes

Una propiedad llamativa de un polinomio cúbico es que tiene un único punto de inflexión. Si se desea construir una curva que tiene más puntos de inflexión, una solución posible es resolver un sistema más grande de ecuaciones lineales. Otra solución posible en este caso sería construir la interpolación uniendo varios polinomios cúbicos. Para ilustrar la opción de usar polinomios cúbicos, supóngase que se desea construir la curva f que se ilustra en en la Figura 4.2.

Para lograrlo se escogen varios puntos llamado comúnmente puntos de control. Sean $(p_1, p_2, \ldots, p_n)$ los puntos de control y $(f_1, f_2, \ldots, f_{n-1})$ las cúbicas que se usarán para interpolar f, tal como se ilustra en la Figura 4.2 con $n = 5$. En el resto de esta sección se asume que $n = 5$, sin que ello implique pérdida de generalidad. Sean (x_i, y_i) las coordenada de p_i. El propósito es ajustar una cúbica entre cada par de puntos de control (p_i, p_{i+1}). Si se denota $f_i(x) = a_{i0} + a_{i1}x + a_{i2}x^2 + a_{i3}x^3$ a la cúbica que une estos dos puntos de control, para ajustar el polinomio 4.3 se requieren 4 ecuaciones para cada cúbica f_i, es decir, $4 \times 4 = 16$ ecuaciones. Para construir estas ecuaciones es necesario imponer algunas condiciones mínimas a las cúbicas f_i que se unirán entre sí para construir el polinomio f.

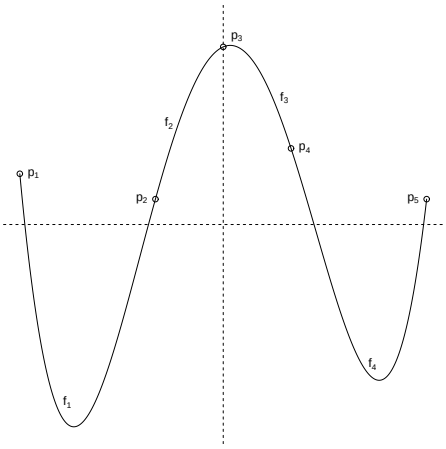


Figura 4.2: Curva que se desea interpolar utilizando los puntos de control $(p, p, \ldots, p)$ y las cúbicas $(f, f, \ldots, f)$.

- *Primera condición: las cúbicas pasan por los puntos de control*
  Es decir que $f_i(x_i) = y_i$ y $f_i(x_{i+1}) = y_{i+1}$, por lo que la cúbica f_i pasa por los puntos p_i y p_{i+1}, generando de esta manera ocho ecuaciones lineales. Como las curvas f_i y f_{i+1} se unen en p_{i+1}, se dice las curvas unidas satisfacen continuidad C^0, indicando que la función que resulte de unir f_i y f_{i+1} es continua y pertenece a un espacio de funciones sin derivadas continuas. En general, se denota C^m al conjunto de todas las funciones continuas que tienen m derivadas continuas.
- *Segunda condición: las cúbicas tienen la misma pendiente en los puntos de control*

Es decir que la primera derivada de f_i coincide con la primera derivada de f_{i+1} en p_{i+1}, con lo que f satisfaría además continuidad C^1. Así que $f_i'(x_{i+1}) = f_{i+1}'(x_{i+1})$. Como $f_i(x) = a_{i0} + a_{i1}x + a_{i2}x^2 + a_{i3}x^3$, se sigue que $f_i'(x) = a_{i1} + 2a_{i2}x + 3a_{i3}x^2$. Esta condición genera tres nuevas ecuaciones lineales, una en cada punto donde se unen dos cúbicas.

- *Tercera condición: las cúbicas tienen la misma curvatura en los puntos de control*

  Lo que implica que $f_i''(x_{i+1}) = f_{i+1}''(x_{i+1})$, garantizando que $f \in C^2$. Como $f_i'(x) = a_{i1} + 2a_{i2}x + 3a_{i3}x^2$, entonces $f_i''(x) = 2a_{i2} + 6a_{i3}x$. Se generan así, como con la segunda condición, tres nuevas ecuaciones lineales, una en cada punto donde se unen dos cúbicas.

- *Cuarta condición (débil): las pendientes de la primera cúbica en p_1 y de la última cúbica en p_5 son conocidas*

  Sean s_1 y s_5 las pendientes de f_1 y f_4 en el primero y en el último puntos de control, respectivamente. Esta condición lo que hace es generar las dos ecuaciones lineales restantes para solucionar el problema: $f_1'(x_1) = s_1$ y $f_4'(x_5) = s_5$. Esta cuarta condición es débil, ya que es poco probable que uno conozca las pendientes inicial y final de la curva que desea interpolar. Así que si se desea utilizar esta curva interpolada para efectivamente encontrar el valor de $f(x)$, lo más prudente sería usarla solamente para valores de X entre p_2 y p_4. Es decir, la interpolación en los extremos del intervalo no es muy precisa.

A manera de ejemplo, sea $A = \{(-3, 2); (-1, 1); (0, 7); (1, 3), (3, 1)\}$ y sea $\pi_4 = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4$ el polinomio que pasa por estos cuatro puntos. Encontremos π_4 usando interpolación por partes.

Usando las condiciones que hemos impuesto, se obtiene la solución 4.5, que se representa en la Figura 4.3.

$$\begin{aligned} f_1 &= 3.7925 - 8.4752x - 15.3895x^2 - 4.1217x^3 \\ f_2 &= 7 + 1.1472x - 5.7670x^2 - 0.9143x^3 \\ f_3 &= 7 + 1.1472x - 5.767x^2 + 0.6198x^3 \\ f_4 &= 3.7840 + 10.7952x - 15.4149x^2 + 3.8357x^3 \end{aligned} \tag{4.5}$$

En esta solución se han utilizado las pendientes $s_1 = -250$ y $s_5 = 230$. El uso de otras pendientes producirá soluciones diferentes

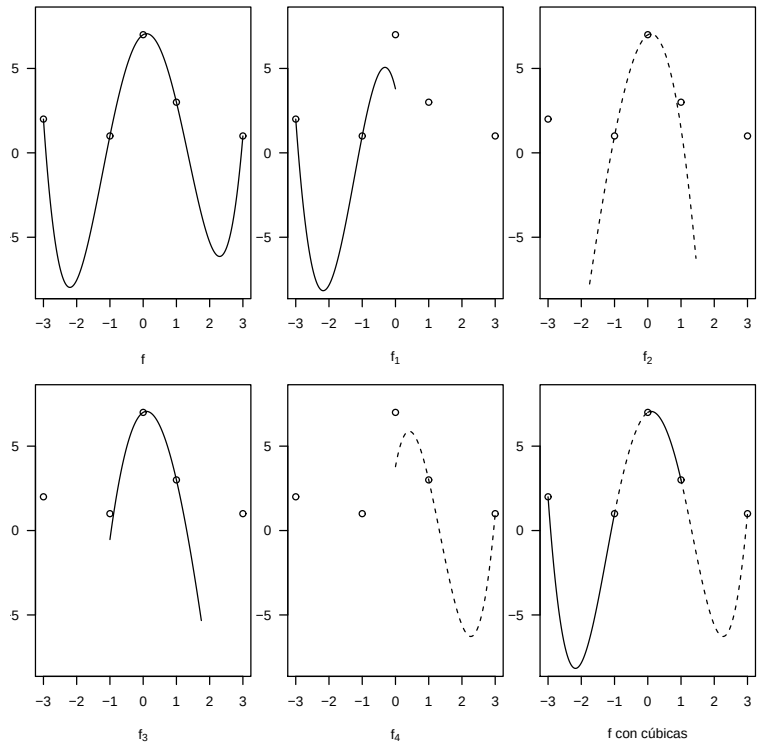


Figura 4.3: Curva a interpolar (arriba a la izquierda), cúbicas utilizadas (f, f, f, f) y curva obtenida (abajo a la derecha).

en el primero y último subintervalos. Esta limitación se debe a que no tenemos información antes del primer punto de control ni después del último.

4.2.3. Estimación

Un problema igualmente interesante desde el punto de vista numérico pero más interesante desde el punto de vista estadístico es aquel en el cual se sabe que los datos disponibles son mediciones sujetas a error y que aunque se cree que la relación entre X y Y es funcional, no necesariamente los puntos observados están sobre la curva que relaciona las dos variables, como se ve en la Figura 4.4. Es posible incluso que se disponga de muchas más observaciones, por ejemplo del tipo ilustrado en la parte derecha de la figura 4.4, en la que se dispone de varios valores observados de Y para cada valor observado de X. Las observaciones del tipo ilustrado en la figura 4.4

provienen de una función desconocida f, que relaciona a X y Y, que se desea estimar. Es decir, el propósito de la estimación es proponer una función $\hat{f}$ que permita aproximar f.

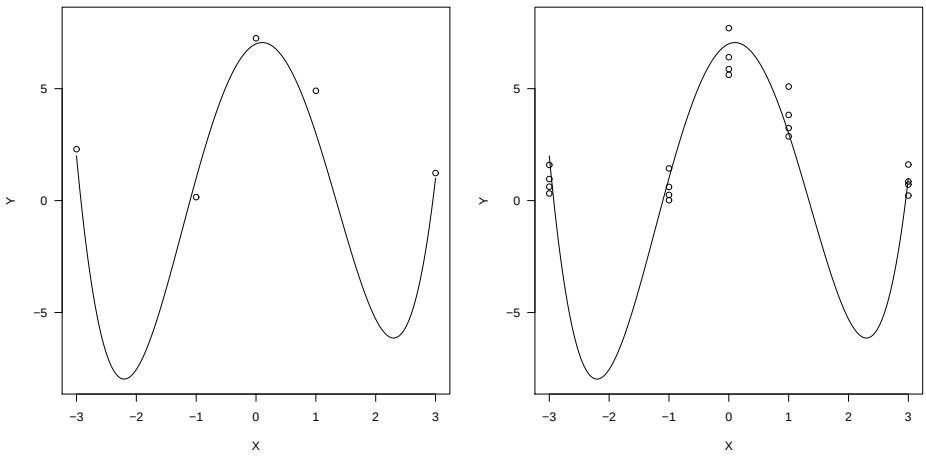


Figura 4.4: Curva que se desea estimar utilizando datos con errores de medición. En el caso de la izquierda se dispone de una observación para cada punto de diseño. En el caso de la derecha, se dispone de varias

En realidad en ambos casos, en la interpolación y en la estimación, se obtienen curvas aproximadas. La diferencia está en que, por una parte, no es posible utilizar interpolación con los puntos de la figura 4.4; y que a través de los métodos de estimación es posible asociar a las curvas construidas una medida probabilística de qué tan buena será la aproximación alcanzada, aun cuando es posible obtener alguna medida de error computacional, no probabilística, en la interpolación.

Una posible solución estadística es la regresión polinómica de la sección 2.4, en la que se asume que la función subyacente f es una función continua que se aproxima mediante un polinomio de Taylor y que las n observaciones disponibles provienen de tal función f, de tal manera que los valores y_i se separan de los de $f(x_i)$ por una cantidad aleatoria ϵ_i, de acuerdo con el modelo 4.1.

Otra solución es la suavización spline, que discutiremos en lo que sigue de esta sección.

4.2.4. Estimación spline

Wahba (1990, pág. viii) y Green & Silverman (2000, pág. 14) describen un spline mecánico como una pieza metálica, plástica,

de madera o de cualquier otro material flexible, que se ajusta a curvas adaptándose a su forma y que permite dibujar curvas *suaves*. Según estos autores, este tipo de herramienta se utilizó en el pasado para delinear cascos de barcos y para planear curvas de carrileras. Así que si fuera posible tener un objeto matemático que actuara como un spline mecánico que tuviera además adecuadas propiedades estadísticas, entonces podría utilizarse para ajustar curvas como las que nos proponemos en esta sección.

La versión más sencilla de un objeto matemático que se comporte como un spline mecánico es llamado un *spline cúbico*. Supongamos que tenemos un conjunto de números reales $x_1, \ldots, x_n$ en un intervalo $[a, b]$, tales que $a < x_1 < x_2 < \ldots < x_n < b$. Una función s definida en $[a, b]$ es un spline cúbico si cumple las siguientes dos condiciones:

1. s es una cúbica en cada uno de los intervalos (a, x_1), (x_1, x_2),...,(x_n, b)

2. La cúbicas se unen en los puntos x_i de tal manera que s y sus dos primeras derivadas con continuas en cada x_i y por lo tanto en todo el intervalo $[a, b]$

A los puntos x_i los llamaremos *nodos*.

Nótese que la función interpolante construida en la sección 4.2.2 satisface la definición de un spline cúbico.

Un spline cúbico en $[a, b]$ se llama un *spline cúbico natural* si se satisface que las dos primeras derivadas de s son iguales a cero en los puntos a y b. Estas condiciones las llamaremos condiciones de acotamiento natural e implican que s es lineal en los dos intervalos extremos (a, x_1) y (x_n, b).

Por otra parte, es posible demostrar (Green & Silverman 2000) que dados unos valores y_i, para un conjunto dado de puntos $x_1 < x_2 < \ldots < x_n$ en $[a, b]$, existe un único spline cúbico natural que satisface que $s(x_i) = y_i, \quad i = 1, \ldots, n$.

Regresando a nuestro problema de estimación, supongamos que deseamos utilizar un spline cúbico natural para estimar f en nuestro modelo 4.1. En la estimación usando series de fourier acordamos que la función de regresión está en $L_2[0, 1]$. Añadiremos ahora la condición de que las dos primeras derivadas de f, que denotaremos f' y f'', también estén en $L_2[0, 1]$. A este espacio de funciones lo llamaremos espacio de Sobolev de orden 2 y lo denotaremos $W_2^2[0, 1]$.

Siguiendo a Eubank (1999) y Green & Silverman (2000), una medida natural de suavidad asociada con una función $f \in W_2^2[0,1]$ es $\int_0^1 f''(x)^2 dx$, mientras que una medida de bondad de ajuste de los datos al modelo es la suma de cuadrados del error $n^{-1}\sum_{i=1}^n (y_i - f(x_i))^2$. Esto implica que una medida de la calidad de un estimador de f podría basarse en la suma convexa:

$$(1-q)n^{-1}\sum_{i=1}^{n}(y_i - f(x_i))^2 + q\int_0^1 f''(x)^2 dx$$

con $0 < q < 1$.

Si hacemos $\lambda = q/(1-q)$, la elección del estimador de f es equivalente a elegir f_λ que minimice la suma:

$$n^{-1}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\int_0^1 f''(x)^2 dx \tag{4.6}$$

sobre todas las funciones $f \in W_2^2[0,1]$. A este estimador f_λ lo llamaremos un estimador spline de f.

De la expresión 4.6 se sigue que si λ es muy grande, entonces la estimación de la función de regresión será super-suavizada; lo contrario ocurre con un λ muy pequeño, que conduce a un estimador que interpola los datos.

Eubank (1999) encuentra que la solución a este problema de optimización es única y corresponde al estimador

$$f_\lambda = \sum_{i=1}^{n}\beta_{\lambda i} f_i \tag{4.7}$$

donde $\boldsymbol{\beta}_\lambda = (\beta_{\lambda 1}, \beta_{\lambda 2}, \ldots, \beta_{\lambda n})^T$ es la única solución con respecto a $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)^T$ del sistema de ecuaciones

$$(\boldsymbol{X}^T\boldsymbol{X} + n\lambda\boldsymbol{\Omega})\boldsymbol{c} = \boldsymbol{X}^T\boldsymbol{y} \tag{4.8}$$

donde

$$\boldsymbol{X} = \{f_j(x_i)\}_{i,j=1,2,\ldots,n}, \quad \boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T \quad \text{y}$$
$$\boldsymbol{\Omega} = \left\{\int_0^1 f_i''(x) f_j''(x) dx\right\}_{i,j=1,2,\ldots,n}.$$

Las funciones $\{f_j\}_{j=1,2,\ldots,n}$ forman una base del conjunto de splines naturales. En particular usaremos la base de splines cúbicos naturales (Sezer 2009)

$$\begin{aligned} f_1(x) &= 1 \\ f_2(x) &= x \\ f_{j+2}(x) &= d_j(x) - d_{n-1}(x), j = 1, 2, \ldots, n-2 \end{aligned} \tag{4.9}$$

donde:

$$d_j(x) = \frac{(x - x_j)_+^3 - (x - x_n)_+^3}{x_j - x_n}$$

y la función $(z)_+^3$ es la función truncada:

$$(z)_+^3 = \begin{cases} z^3, & \text{si } z \geq 0 \\ 0, & \text{si } z < 0 \end{cases}$$

En consecuencia, el vector de valores estimados es $\boldsymbol{f}_\lambda = (f_\lambda(x_1), f_\lambda(x_2), \ldots, f_\lambda(x_n))^T = \boldsymbol{S}_\lambda \boldsymbol{y}$, donde tenemos que

$$\boldsymbol{S}_\lambda = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X}^T \tag{4.10}$$

Al estimador f_λ de f definido en 4.7 le llamaremos un *estimador spline*. La elección del parámetro de suavización λ se hace usualmente con el estimador de validación cruzada generalizada GCV, usando la matriz $\boldsymbol{S}_\lambda$ 4.10.

La figura 4.5 (izquierda) muestra la estimación spline de los datos de CO en Cali; en la parte derecha se compara la estimación spline con las otras dos estimaciones propuestas, por series de cosenos y kernel (Nadaraya-Watson). El valor de lambda se selecciona por validación cruzada y la estimación se conduce utilizando la función *smooth.spline* del paquete *stats* (R Development Core Team 2011) de **R**.

4.3. Estimación spline por mínimos cuadrados

Las propiedades asintóticas de los estimadores spline que hemos discutido hasta aquí se derivan del hecho de que estos estimadores pueden representarse como estimadores kernel (Simonoff 1996), de tal manera que sus propiedades asintóticas son similares. De hecho, los estimadores splines convergen a la función de regresión a una tasa de

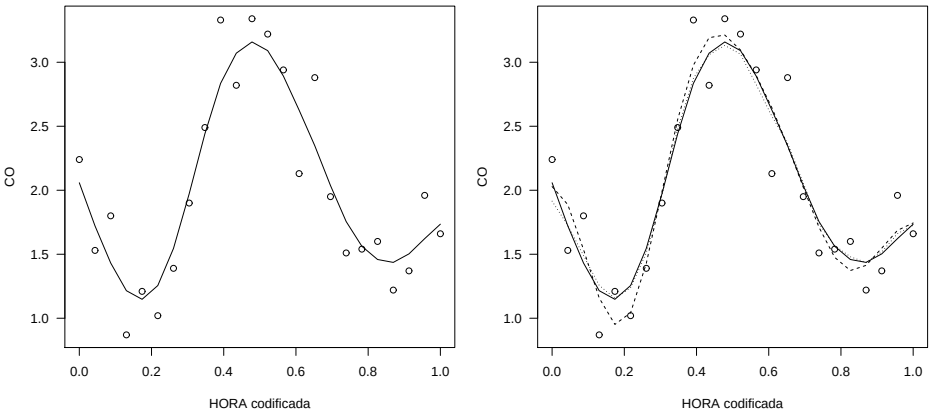


Figura 4.5: Estimación spline de la contaminación por CO en Cali. En la derecha, se compara este estimador con los estimadores óptimos de series de cosenos y kernel (Nadaraya-Watson). Los estimadores spline (línea continua) y kernel (línea punteada) casi se superponen, mientras que el estimador de series (línea segmentada) parece sub-estimar en los extremos y sobre-estimar en el centro del diseño (por lo menos con respecto a los otros dos estimadores).

convergencia óptima de $n^{4/5}$, precisamente por sus similaridades con los estimadores kernel.

Existe sin embargo otro posible estimador spline que no puede representarse como un estimador kernel. Esta nueva forma de estimación se apoya en el modelo de regresión polinómica discutido en la sección 2.4. En ese caso representamos el modelo 4.1 de la siguiente manera:

$$f(x_i) = \sum_{j=0}^{\lambda} \theta_j f_j(x_i) + r_\lambda(x_i), \quad i = 1, \ldots, n \tag{4.11}$$

donde λ cumple en tal caso el papel de parámetro de suavización y consiste en el número de términos que se conservan en la sumatoria.

Si modificamos la notación para llamar m a lo que en la expresión 4.11 hemos denotado como λ, una de varias representaciones posibles para $r_m(x)$ sería:

$$r_m(x) = \frac{1}{m-1} \int_0^1 f^{(m)}(x)(x-\xi)_+^{m-1} d\xi \tag{4.12}$$

Si $r(x_1), \ldots, r(x_n)$ son muy pequeños, entonces el modelo de regresión polinómica 2.50 sería adecuado para un conjunto particular de datos que tenga estas características. Pero si este no es el caso, una posible generalización es la estimación spline por mínimos cuadrados, que se apoya en la siguiente aproximación a la integral en la expresión

4.12:

$$r_m(x) = \sum_{j=1}^{k} \delta_j (x - \xi_j)_+^{m-1} \tag{4.13}$$

para un conjunto de constantes $\delta_1, \ldots, \delta_k$ y un conjunto de puntos $0 < \xi_i < \ldots < \xi_k < 1$, de tal manera que una nueva aproximación a la función de regresión lucirá ahora así:

$$f(x) = \sum_{j=1}^{\lambda} \theta_j x^{j-1} + \sum_{j=1}^{k} \delta_j (x - \xi_j)_+^{m-1} \tag{4.14}$$

Sea λ un conjunto dado de puntos $\{\xi_i < \ldots < \xi_k\}$, una posible solución para estimar f sería estimar por mínimos cuadrados los m coeficientes $\{\theta_j\}_{j=1}^{m}$ y los k coeficientes $\{\delta_j\}_{j=1}^{k}$. Sea $\beta = (\theta_1, \ldots, \theta_m, \delta_1, \ldots, \delta_k)^T$ el vector de coeficientes del modelo. Entonces el estimador spline de mínimos cuadrados de f será:

$$f_\lambda(x) = \sum_{j=1}^{m+k} \beta_{\lambda j} f_j(x) \tag{4.15}$$

Un problema aún en curso es la elección de m y de k. A los k valores de ξ se les llama nodos y la solución más sencilla es hacer uso de inspección visual de los datos, cuando sea posible. Si f cambia muy rápidamente, entonces deberemos usar más nodos. La elección de m depende de si se usará un spline lineal, cuadrático, cúbico, etc. y los k nodos se ubicarán consecuentemente.

Nuevos desarrollos recientes (Fromkorth & Kohler 2011) muestran la importancia de esta técnica en muchas aplicaciones, como el caso de la determinación de precios de opciones. Ver además Kohler & Krzyúzak (2010)

4.4. Ejercicios

1. La biblioteca Dierckxspline del software de procesamiento estadístico de libre distribución **R** incluye un conjunto de datos (titanium) sobre una propiedad no determinada del Titanio como función del calor. Ajuste un modelo spline por mínimos cuadrados usando el conjunto de $k = 10$ puntos equidistantes entre 800 y 1100.

2. Estime f en el problema del Titanio 1 usando un estimador kernel y un estimador spline y compare sus estimaciones con la obtenida por spline por mínimos cuadrados.

3. Escriba un programa en **R** para estimar la función de regresión usando un estimador spline de mínimos cuadrados. Utilice su programa para estimar el comportamiento diario de la contaminación por CO en el centro de Cali (Ejercicio 2.1).

PÁGINA EN BLANCO
EN LA EDICIÓN IMPRESA

Capítulo 5

MODELOS ADITIVOS GENERALIZADOS

5.1. Introducción

Para establecer un marco general, asumiremos que disponemos de observaciones de la variable de respuesta Y para n valores predeterminados de una variable independiente X. Las n observaciones bivariadas disponibles, denotadas $(x_1, y_1), \ldots, (x_n, y_n)$, siguen el modelo

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{5.1}$$

donde $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ es un vector de errores aleatorios no correlacionados que tienen media cero y varianza común σ^2, f es una función de regresión desconocida y se satisface que $0 \leq x_1 < \cdots < x_n \leq 1$.

Para efectos de presentación de los resultados, asumiremos que los valores de X han sido elegidos así:

$$x_i = (2i - 1)/2n, \quad i = 1, 2, \ldots, n \tag{5.2}$$

Nuestro propósito es estimar f en 5.1, para lo cual buscaremos construir estimadores lineales que puedan escribirse en la siguiente forma general, que para un λ dado es una combinación lineal de las observaciones y_i, donde $K(\cdot, x_i; \lambda)$, $i = 1, \ldots, n$ es una colección de funciones ponderadoras que dependen de los x_i y de un parámetro de

suavización denotado λ:

$$f_\lambda(x) = \sum_{i=1}^{n} K(x, x_i; \lambda) y_i \tag{5.3}$$

5.2. Modelos Aditivos Generalizados GAM

Para resolver el problema de la maldición de la dimensionalidad en la estimación multivariante se han propuesto varias soluciones, todas encaminadas a reducir la dimensión en la cual se conduce el análisis de regresión.

Un ejemplo de una manera eficiente de reducir la dimensionalidad es SIR (Sliced Inverse Regression) (Li 1991). SIR reduce la dimensión conservando $k < p$ combinaciones lineales de todos las p variables explicativas. Es decir, tal como ya se mencionó, se reduce la dimensión, pero se conservan todas las variables originales.

Otro procedimiento para reducir la dimensionalidad es PPR (Projection Pursuit Regression) (Friedman & Stuetzle 1981, Hall 1989). En este caso también se conservan k combinaciones lineales de las p variables explicativas, con $k < p$. Los dos métodos difieren en la forma en que cada uno escoge las combinaciones lineales y en la forma como ajustan el modelo final.

SIM (Single Index Models) (Härdle, Hall & Ichimura 1993) es un ejemplo más. En este caso se conserva una y solo una combinación lineal de las p variables explicativas originales y se ajusta un modelo univariado. SIM obtiene la combinación lineal de la misma forma que PPR, pero las dos ideas inician desde suposiciones diferentes.

El método Π de Breiman (1991) estima una función de regresión en p variables como una suma de k (de nuevo $k < p$) productos de funciones suaves univariadas de las variables explicativas. Un ajuste de este tipo luce así:

$$\sum_{i=1}^{k} \prod_{j=1}^{p} \phi_{i,j}(X_j).$$

Aquí las funciones $\phi_{i,j}(X_j)$ se obtienen por suavización unidimensional. De esta manera, el modelo final requiere k productos de p funciones unidimensionales, pero no se requiere suavización multidimensional en ninguna etapa. Los resultados finales son difíciles de interpretar, lo que hace el método de poca utilidad práctica.

Pero todos estos esfuerzos se han reducido a la propuesta de solución más aceptada hasta ahora para manejar el problema de la dimensionalidad, que son los Modelos Aditivos Generalizados GAM (Generalized Additive Models) (Hastie & Tibshirani 1990). En este caso la función de regresión se representa como una suma de funciones univariadas sencillas, una por cada variable explicativa. La función de regresión luce así:

$$\sum_{i=1}^{p} g_i(X_i) \tag{5.4}$$

Las funciones unidimensionales g_i se obtienen a través de un procedimiento iterativo llamado *backfitting* (Hastie & Tibshirani 1990). El backfitting se puede describir de la siguiente manera. Supóngase que la función de regresión puede expresarse como la suma de dos funciones unidimensionales, por ejemplo: $g(\mathbf{X}) = g_1(X_1) + g_2(X_2)$. Se requiere entonces ajustar un modelo de la forma:

$$Y = g_1(X_1) + g_2(X_2) + \epsilon, \tag{5.5}$$

para un error aleatorio ϵ que tiene media 0 y varianza finita.

Si el modelo (5.5) es correcto, entonces $E[Y - g_i(X_i)|X_j] = g_j(X_j)$, $i, j = 1, 2$. Se sigue que, dada una estimación $\hat{g}_1(x_1)$ de $g_1(x_1)$, una manera intuitiva de estimar $g_2(x_2)$ sería suavizar los residuales $Y - \hat{g}_1(x_1)$ sobre X_2. Entonces, dada una estimación $\hat{g}_1(x_1)$, una manera intuitiva de estimar $g_2(x_2)$ sería suavizar los residuales $Y - \hat{g}_1(x_1)$ sobre X_2. Con esta estimación $\hat{g}_2(x_2)$ es posible obtener una estimación mejorada de $\hat{g}_1(x_1)$ suavizando $Y - \hat{g}_2(x_2)$ sobre X_1. El proceso continúa hasta que:

1 La estimación $\hat{g}_1(x_1)$ resultante de suavizar $Y - \hat{g}_2(x_2)$ sobre X_1 es muy cercana a la estimación $\hat{g}_1(x_1)$ utilizada para estimar $\hat{g}_2(x_2)$ en la iteración previa, y

2 La estimación $\hat{g}_2(x_2)$ resultante de suavizar $Y - \hat{g}_1(x_1)$ sobre X_2 es muy cercana a la estimación $\hat{g}_2(x_2)$ utilizada para estimar $\hat{g}_1(x_1)$ en la iteración previa.

En este momento se escoge $\hat{g}_1$ como estimador de g_1 y $\hat{g}_2$ como estimador de g_2. Si es posible expresar la función de regresión g como una suma del tipo (5.4), entonces se esperaría conservar una función g_i unidimensional para cada variable explicativa. De esta manera, GAM

se hace cargo del problema de la suavización p-dimensional, y utiliza todas las p variables explicativas en el modelo final.

La figura 5.1 muestra un ajuste GAM a los datos del problema de la contaminación por Ozono en Rio de Janeiro. En este caso, el modelo que nos proponemos ajustar tiene la forma general:

$$Y = g_1(X_1) + g_2(X_2) + \epsilon, \tag{5.6}$$

donde las funciones g_1 y g_2 resultan por suavización kernel unidimensional, en esta caso usando estimadores kernel lineales locales. Se observa que la temperatura incide mucho más en los niveles de Ozono troposférico que la humedad relativa. Los valores de la respuesta son centrados.

Un modelo de regresión lineal bivariada se escribe generalmente como:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_1 x_{2i} + \epsilon_i, \tag{5.7}$$

por lo que un modelo GAM bien podría escribirse así:

$$y_i = \beta_0 + g_1(x_{1i}) + g_2(x_{2i}) + \epsilon_i, \tag{5.8}$$

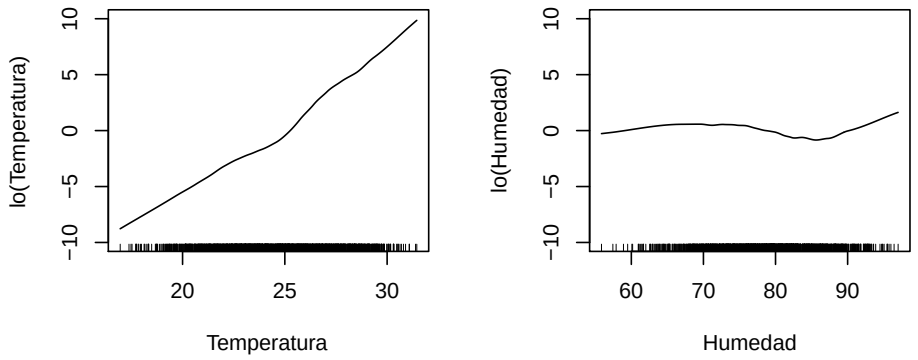


Figura 5.1: Estimación GAM del comportamiento del Ozono troposférico (promedio diario), dependiendo de la temperatura y la humedad promedio diarias en Rio de Janeiro en el periodo 2001-2005, usando la función *gam* con suavización local de **R**.

Para estimar β_0 en el modelo 5.8 podríamos utilizar técnicas paramétricas, lo que convertiría este modelo en un modelo semi-paramétrico que incluso podría lucir, aún más generalmente, como:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_1 x_{2i} + g_1(x_{1i}) + g_2(x_{2i}) + \epsilon_i, \tag{5.9}$$

Y para el caso de p variables, sería:

$$y_i = \boldsymbol{x}_i^T \beta + \sum_{j=1}^{p} g_j(x_{ji}) + \epsilon_i, \quad i = 1, \ldots, n \tag{5.10}$$

Pero el gran aporte de los modelos aditivos generalizados GAM no está en la eliminación de la necesidad de la suavización múltiple, con sus dificultades dimensionales. El aporte esencial está en que, en efecto, permiten *generalizar* los modelos de regresión no paramétrica tal como los modelos lineales generalizados introducidos por Nelder & Wedderburn (1972) y popularizados por Nelder & McCoullagh (1989) cumplen con su papel de *generalizar* los modelos lineales.

5.3. Modelos aditivos generalizados: regresión logística no paramétrica

Ya anticipamos que el gran valor de los modelos GAM de Hastie & Tibshirani (1990) es su asociación con los modelos lineales generalizados de Nelder & Wedderburn (1972). Para verlo, en Green & Silverman (2000) se presenta una propuesta de unificación de los modelos de regresión a través del mecanismo de separar el modelo 5.1 en dos componentes, una aleatoria y una sistemática. Se introduce entonces un vector de predictores θ_i, uno por cada observación, de tal manera que la componente aleatoria permita especificar la forma en la que la distribución de Y_i depende de θ_i. Al mismo tiempo, la componente sistemática permite especificar la estructura de los θ_i como función de las variables explicativas.

Por ejemplo, en los modelos lineales clásicos la componente aleatoria se distribuye Normal con media θ_i y varianza σ^2, mientras que la componente sistemática sería $\theta_i = \boldsymbol{x}_i^T \beta,\ i = 1, 2, \ldots, n$. En el caso de la regresión no paramétrica univariada, la componente sistemática podría escribirse como $\theta_i = f(x_i),\ i = 1, 2, \ldots, n$. Y más aún, para un modelo semiparamétrico, esta componente sistemática sería $\theta_i = \boldsymbol{x}_i^T \beta + f(x_i),\ i = 1, 2, \ldots, n$.

El ejemplo más sencillo de un modelo lineal generalizado no paramétrico en el modelo de regresión logística no paramétrica que nos proponemos describir en esta sección. Supongamos que nos interesa ajustar un modelo de regresión logística a una variable de respuesta dicotómica Y_i que observamos en puntos x_i. Este modelo paramétrico

tomaría la forma $\text{logit}[P(Y_i = 1)] = \beta_0 + \beta_1 x_i$ y nuestro interés sería estimar β_0 y β_1. El modelo de regresión logística no paramétrica reemplaza la dependencia de la respuesta de una recta, reemplazando esta recta por una función *suave* f, de tal manera que el nuevo modelo sería $\text{logit}[P(Y_i = 1)] = f(x_i)$. Además, la componente aleatoria ya no es Normal si no, en este caso, Bernoulli. Es decir, Y_i se distribuye Bernoulli con parámetro θ_i y la componente sistemática sería $\text{logit}[\theta_i] = f(x_i)$.

Nos proponemos tratar modelos como este usando inicialmente un modelo lineal generalizado (GLM) no paramétrico y luego un modelo aditivo generalizado (GAM).

En los GLM se asume que las respuestas y_i son elegidas de manera independiente de una familia exponencial de un solo parámetro. La función de densidad o de probabilidad de Y_i podría describirse en general como:

$$p(y_i; \theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right) \tag{5.11}$$

donde θ_i es el parámetro de la distribución relativo a Y_i, que se espera recoja información de las variables explicativas; ϕ es un parámetro de perturbación o de escala común a los Y_i, cuyo papel es similar al de σ^2 en los modelos lineales paramétricos; y, finalmente, las funciones a y b determinan la forma específica de la distribución.

Para determinar la parte sistemática del modelo se define la forma funcional de θ_i en términos de las variables explicativas para la respuesta i. Sea μ_i el valor esperado de $(Y_i; \theta_i, \phi)$. Entonces para la distribución 5.11 tenemos que $\mu_i = E[Y_i; \theta_i, \phi] = b'(\theta_i)$.

Los GLM se construyen asumiendo que existe una *función de enlace* G tal que $G(\mu_i) = \boldsymbol{x}_i^T\beta$. Para el caso de la regresión logística se tiene, con m_i el denominador binomial (número de casos totales en cada x_i), que:

$$\begin{aligned} b(\theta_i) &= m_i \log(1 + e^{\theta_i}) \\ \phi &= 1 \\ c(y_i, \phi) &= -\log\binom{m_i}{y_i} \\ G(b'(\theta_i)) &= \boldsymbol{x}_i^T\beta = \log\frac{\mu_i}{m_i - \mu_i} \end{aligned}$$

En un GLM no paramétrico sustituimos $G(b'(\theta_i)) = \boldsymbol{x}_i^T\beta$ por la expresión más general $\theta_i = f(x_i)$, de tal manera que buscaremos una función que maximice la suma:

$$l(f) - \lambda \int_0^1 f''(x)^2 dx$$

donde $l(f)$ es el logaritmo de la verosimilitud y λ es el parámetro que controla la suavidad del ajuste.

En un GAM buscaremos f_λ que maximice:

$$\Pi = l(f) - \sum_{j=1}^{p} \lambda_j \int_0^1 f_j''(x)^2 dx$$

Utilizaremos un GAM para ajustar un modelo de regresión logística al problema de la contaminación por Ozono troposférico en Rio de Janeiro. El propósito es ajustar un modelo que nos permita estimar la probabilidad de que se produzca una contaminación de Ozono por encima de 15ppm de O_3, dependiendo de la temperatura promedio diaria. La figura 5.2 muestra un modelo logístico paramétrico (línea punteada) y un modelo logístico no paramétrico (linea continua). El modelo paramétrico parece subestimar las probabilidades de ocurrencia de este fenómeno para temperaturas altas y bajas. A temperaturas medias los dos modelos se comportan de manera muy similar.

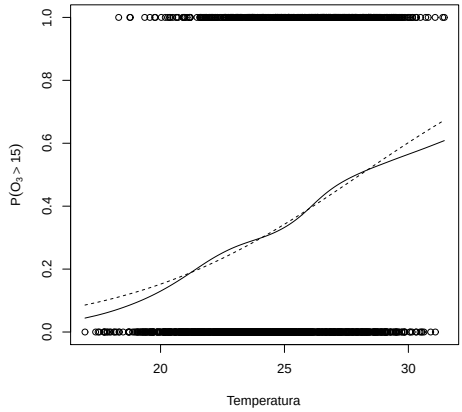


Figura 5.2: Ajuste de un modelo logístico paramétrico (línea punteada) y no paramétrico (linea continua) a los datos de contaminación por Ozono en Rio de Janeiro

5.4. Ejercicios

1. Utilice los datos del famoso experimento de Hald (Draper & Smith 1998, Pág. 348) y ajuste un modelo aditivo generalizado GAM usando las dos variables más importantes como variables de predicción. Compare este modelo con el mejor modelo lineal.

2. En lo que podríamos llamar un ejemplo clásico, Hastie & Tibshirani (1990, Pág. 301) usan el conjunto de datos kyphosis de la biblioteca gam del software de procesamiento estadístico de libre distribución **R** para ilustrar el uso de los modelos aditivos generalizados. Ajuste un GAM usando la variable kyphosis como respuesta y las variables start y number como variables de predicción. Ajuste un modelo logístico con las mismas variables y compare sus resultados.

Capítulo 6

RESPUESTAS MÚLTIPLES

6.1. Introducción

En el análisis de regresión se asume en general que disponemos de observaciones de la variable de respuesta Y para n valores predeterminados de una variable independiente X. Las n observaciones bivariadas disponibles, denotadas $(x_1, y_1), \ldots, (x_n, y_n)$, siguen el modelo

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{6.1}$$

donde $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ es un vector de errores aleatorios no correlacionados que tienen media cero y varianza común σ^2 y f es una función de regresión desconocida. Asumiremos que $0 \leq x_1 < \cdots < x_n \leq 1$.

Para efectos de presentación de los resultados teóricos, se acostumbra además que los valores de X se elijan de la siguiente manera:

$$x_i = (2i - 1)/2n, \quad i = 1, 2, \ldots, n \tag{6.2}$$

Si nuestro propósito es estimar f en el Modelo (6.1), buscaremos construir estimadores lineales que puedan escribirse en la siguiente forma general, que para un λ dado (al que llamaremos "parámetro de suavización") es una combinación lineal de las observaciones y_i, donde $K(\cdot, x_i; \lambda),\ i = 1, \ldots, n$ es una colección de funciones ponderadoras

que dependen de los x_i y de λ:

$$f_\lambda(x) = \sum_{i=1}^{n} K(x, x_i; \lambda) y_i \tag{6.3}$$

En la Ecuación (6.3) la función K es una función simétrica y centrada en cero que tiene su máximo en cero. A estas funciones se les llama funciones kernel y los estimadores lineales construidos con estas funciones se les llama estimadores kernel.

Sin embargo, en algunos problemas se dispone de más de una respuesta para cada valor de X. Aunque hay muchas versiones de notación para este caso, nos apoyaremos en la propuesta por Draper & Smith (1998, pág. 49). Supongamos en particular que tenemos m valores diferentes de X y que se dispone de $n_j, j = 1, 2, \ldots, m$ respuestas para cada x_j, tales que $\sum_{j=1}^{m} n_j = n$. En tal caso nuestro Modelo (6.1) podría re-escribirse como:

$$y_{ju} = f(x_j) + \epsilon_{ju}, \quad u = 1, \ldots, n_j, \quad j = 1, \ldots, m \tag{6.4}$$

En este caso nuestros valores de X serían elegidos así:

$$x_j = (2j - 1)/2m, \quad j = 1, 2, \ldots, m \tag{6.5}$$

Obsérvese que esta notación tiene el valor agregado de reservar el contador i para las mediciones individuales, que varían desde 1 hasta n.

Nuestro propósito es estimar f en el Modelo (6.4), para lo cual buscaremos construir estimadores lineales que puedan escribirse en la siguiente forma general, que para un λ dado es una combinación lineal de las observaciones y_{ju}, donde $K(\cdot, x_j; \lambda),\ j = 1, \ldots, m$ es una colección de funciones ponderadoras que dependen de los x_j y de λ:

$$f_\lambda(x) = \sum_{j=1}^{m} \sum_{u=1}^{n_j} K(x, x_j; \lambda) y_{ju} \tag{6.6}$$

Esto significa que asignaremos el mismo peso en la estimación de f a todos los y_{ju} asociados con x_j, para cada j.

Ejemplo 6.1 (***NO_2 en el centro de Cali***)**.** La Red de Vigilancia (Monitoreo) de la Calidad de Aire (RMCA) de Cali colecta información sobre varios contaminantes atmosféricos entre los cuales se

encuentra el Dióxido de Nitrógeno NO_2, un contaminante primario que entre otros efectos actúa como precursor de Ozono (O_3) troposférico. Los equipos de la RMCA miden cada 10 segundos el nivel de NO_2 en el aire, pero reportan el promedio horario (conocido generalmente como "contaminación 1H"), y el mínimo y el máximo en una hora. De esta manera, cada día se dispone de un máximo de 24 observaciones, una para cada hora del día. La Figura 6.1, izquierda, muestra los niveles 1H de NO_2 entre los días miércoles 20 y domingo 24 de enero de 2004 en una estación en el centro de Cali. Las observaciones de cada día se han separado con una línea vertical discontinua.

Un primera mirada a la representación de la Figura 6.1 (izquierda) parece indicar que en cada día se produce un máximo en algún momento alrededor del mediodía y que los días entre semana (miércoles, jueves y viernes) tienen mínimos más bajos y máximos más altos que los días de fin de semana (sábado y domingo).

Se desea encontrar el comportamiento diario del contaminante NO_2 con fines de estimación y, eventualmente, de pronóstico. Para efectos de pronóstico una primera aproximación analítica ante un conjunto de datos como este sería utilizar un análisis de series temporales. Algunos análisis previos con datos similares no han sido muy exitosos en su propósito de ajustar un modelo eficiente de series de tiempo (Barrientos, Olaya & Gonzalez 2007), aunque desde luego hay aún mucho por trabajar en este campo, incluida una aproximación no paramétrica a las series de tiempo.

En cuanto al estudio del comportamiento diario, una posibilidad analítica sería considerar cada día como un "individuo" que se observa a lo largo del tiempo, como en los casos de análisis de datos longitudinales (Diggle et al. 2002), o pensar los datos como datos funcionales que se asumen como realizaciones de un proceso aleatorio con curvas suaves $f(t)$ que se observan en tiempos discretos (Wang 2003). En este trabajo exploraremos el uso de modelos clásicos de regresión no paramétrica (Takezawa 2006, Eubank 1999, Bowman & Azzalini 1997, Fan & Gijbels 1996, Simonoff 1996) bajo el esquema de modelos con respuestas múltiples por punto de diseño. Consideremos la reorganización de los datos tal como se representan en la Figura 6.1, derecha, en la que las observaciones de la hora j en cada uno de los días se han representado como respuestas múltiples en la hora j de cada día. Todas las figuras y los cálculos en este trabajo se realizan con el lenguaje de libre distribución **R** (R Development Core Team 2011).

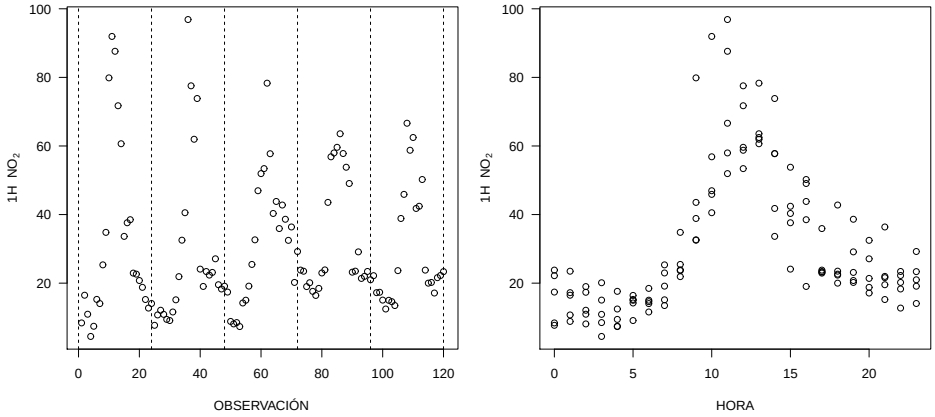


Figura 6.1: Niveles 1H de NO entre los días miércoles 20 y domingo 24 de enero de 2004 en una estación en el centro de Cali. Arriba los datos tomados cronológicamente. Abajo los datos representados como medidas repetidas.

Como ya hemos anotado, los días ordinarios lucen diferentes de los días de fin de semana (y podemos añadir en este grupo los días festivos). En la Figura 6.2 hemos representado a la izquierda los días de miércoles a viernes y a la derecha los días sábado y domingo. Puede observarse que el comportamiento general, sin acudir a ninguna estrategia de suavización o de regresión, parece ser definitivamente diferente para estos tipos de días. En general separaremos el análisis para los días entre semana (días ordinarios) y para los fines de semana y festivos (días festivos).

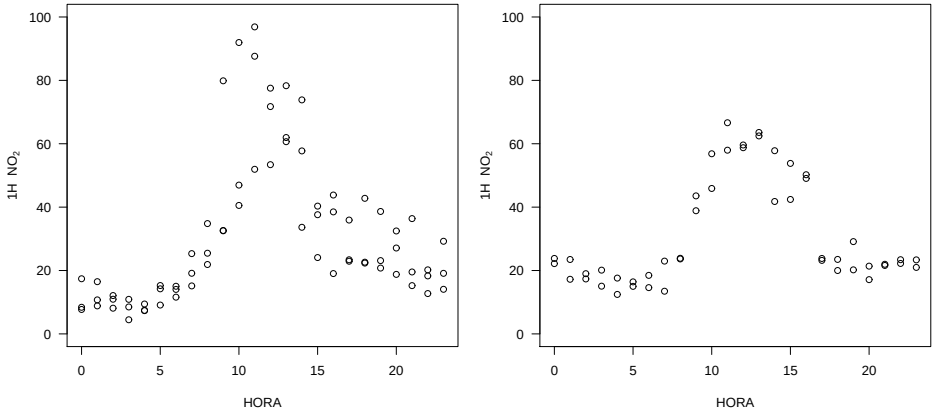


Figura 6.2: Niveles 1H de NO entre los días miércoles 20 y domingo 24 de enero de 2004 en una estación en el centro de Cali. Arriba los datos de días entre semana. Abajo los datos de días de fin de semana.

□

En el caso de la regresión no paramétrica con n valores x_i diferentes y una respuesta y_i asociada con cada x_i, se dispone de un buen número soluciones en la literatura, incluso cuando la respuesta no es continua (Green & Silverman 2000). Pero en este caso nos interesa estimar la función de regresión en el Ejemplo 6.1, que no encaja en la forma general del Modelo (6.1), debido a que disponemos de más de una respuesta para cada punto de diseño. En el ambiente de los modelos lineales paramétricos, este problema está resuelto desde hace muchos años (Draper & Smith 1966). Pero en el caso de la regresión no paramétrica el problema es aún fuente de discusión.

Funciones suaves

Asumiremos que la función f del Modelo (6.1) es una función cuadrado integrable (es decir definida en el espacio $L_2[0,1]$) que tiene dos derivadas continuas. Esta colección infinita de funciones cuadrado integrables con dos derivadas continuas forma un espacio funcional al que denotaremos $W_2^2[0,1]$. A las funciones del espacio $W_2^2[0,1]$ las llamaremos funciones "*suaves*" y al proceso para encontrar una función f en el Modelo (6.1) lo llamaremos "*suavización*".

Supongamos ahora que existe una base de funciones $\{f_j\}_{j=1}^{\infty}$ que permite generar el espacio $W_2^2[0,1]$ y un conjunto de coeficientes $\{\beta_j\}_{j=1}^{\infty}$ tales que la función f puede representarse usando la expansión

$$f = \sum_{j=1}^{\infty} \beta_j f_j \tag{6.7}$$

En tal caso, el Modelo (6.1) puede representarse como:

$$y_i = \sum_{j=1}^{\infty} \beta_j f_j(x_i) + \epsilon_i,\ i = 1, 2, \ldots, n \tag{6.8}$$

lo que significa que los datos siguen un modelo lineal con infinitos coeficientes de regresión desconocidos.

Si los β_j decayeran a cero consistentemente a medida que se usan más de ellos para representar f, entonces uno podría asumir que existe un entero λ tal que

$$f \doteq \sum_{j=1}^{\lambda} \beta_j f_j$$

y por tanto que podríamos escribir la aproximación

$$y_i \doteq \sum_{j=1}^{\lambda} \beta_j f_j(x_i) + \epsilon_i, \;\; i = 1, 2, \ldots, n \tag{6.9}$$

Pero este Modelo (6.9) luce tal como un modelo lineal, por lo que una posible solución al problema de la estimación de f sería estimar los coeficientes $\{\beta_j\}_{j=1}^{\lambda}$ usando el método de mínimos cuadrados, para lo cual definiremos la matriz $\boldsymbol{X}_\lambda = \{f_j(x_i)\}_{i=1,2,\ldots,n;\; j=1,2,\ldots,\lambda}$. Se sigue que el estimador de β tendrá la forma general:

$$\boldsymbol{\beta}_\lambda = (\boldsymbol{X}_\lambda^T \boldsymbol{X}_\lambda)^{-1} \boldsymbol{X}_\lambda^T \boldsymbol{y} \tag{6.10}$$

con $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$.

Entonces nuestro estimador de f será:

$$f_\lambda(x) = \sum_{j=1}^{\lambda} \beta_{\lambda j} f_j(x) \tag{6.11}$$

que puede escribirse como

$$\boldsymbol{f}_\lambda = \boldsymbol{S}_\lambda \boldsymbol{y} \tag{6.12}$$

con $\boldsymbol{f}_\lambda = (f_{\lambda 1}, f_{\lambda 1}, \ldots, f_{\lambda n})^T$ y $\boldsymbol{S}_\lambda = \boldsymbol{X}_\lambda (\boldsymbol{X}_\lambda^T \boldsymbol{X}_\lambda)^{-1} \boldsymbol{X}_\lambda^T$. La matriz $\boldsymbol{S}_\lambda$ luce tal como la matriz "hat" ($\boldsymbol{H}$) en los modelos lineales y juega el mismo papel en la estimación de la función de regresión.

Pero esta solución, que se usa para el Modelo (6.1) no parece adecuada para el Modelo (6.4) que se adecúa más al Ejemplo 6.1.

6.2. La aproximación de Bowman y Azzalini

Para ajustar una curva suave al conjunto de datos del Ejemplo 6.1, Bowman & Azzalini (1997, Pág. 137) los visualizan como un conjunto de "perfiles" de N individuos. En nuestro caso los individuos serán los días (miércoles 20 de enero de 2004, jueves 21 de enero de 2004, etc.) y usando la notación de Draper & Smith (1998), tendremos n_j mediciones a la hora x_j para cada uno de los N días, con $n_j \leq N$. La Figura 6.3 muestra los perfiles para los días laborales (miércoles a viernes) del Ejemplo 6.1.

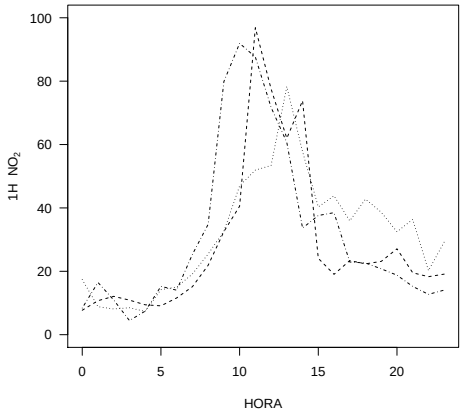


Figura 6.3: Perfiles de la contaminación 1H de NO para los días miércoles 20, jueves 21 y viernes 22 de enero de 2004 en el centro de Cali

Para estimar f, Bowman & Azzalini (1997) proponen ajustar un modelo del tipo (6.1) usando como única respuesta para cada j la media $\bar{y}_j$ de las n_j mediciones asociadas con la hora x_j, es decir,

$$\bar{y}_j = \frac{1}{n_j} \sum_{u=1}^{n_j} y_{ju}. \tag{6.13}$$

La función de regresión se estima de manera habitual usando un estimador lineal de la forma

$$\boldsymbol{f}_\lambda = \boldsymbol{S}_\lambda \bar{\boldsymbol{y}} \tag{6.14}$$

con $\bar{\boldsymbol{y}} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m)^T$ y $\boldsymbol{S}_\lambda$ resultante de usar las funciones de pesos de la Ecuación (6.6) para un parámetro de suavización λ dado.

En esta solución los autores asumen que los datos colectados en diferentes días pueden considerarse independientes, mientas que los datos provenientes de cualquier día individual podrían estar correlacionados. Si además la covarianza es estacionaria, entonces la estructura de dependencia de los ϵ_{ju} tiene la forma general

$$cov\{y_{ju}, y_{kh}\} = cov\{\epsilon_{ju}, \epsilon_{kh}\} = \begin{cases} \sigma^2 \rho_{|u-h|}, & \text{si } j = k \\ 0, & \text{si } j \neq k \end{cases} \tag{6.15}$$

donde σ^2 es la varianza del proceso y $\rho_0 = 1$.

Si denotamos V la matriz $m \times m$ de covarianzas de cada perfil y_j, entonces sus entradas serán:

$$cov\{y_{ju}, y_{kh}\} = V_{ju} = \sigma^2 \rho_{|u-h|} \tag{6.16}$$

En este contexto, a diferencia de los modelos de series de tiempo, la función de autocorrelación $\{\rho_1, \rho_1, \ldots\}$ se considera una componente de perturbación, por lo que no se modela. De hecho, la operación de promediar los y_{ju} para cada x_j preserva la función de autocorrelación por cuanto:

$$cov\{\bar{y}_u, \bar{y}_h\} = \frac{1}{n_j}\sigma^2\rho_{|u-h|} \tag{6.17}$$

Los autores señalan que la estructura de correlación afecta la varianza del estimador, pero no la media, por cuanto:

$$E[\boldsymbol{f}_\lambda] = \boldsymbol{S}_\lambda \boldsymbol{f} \qquad\qquad Var[\boldsymbol{f}_\lambda] = N^{-1}\boldsymbol{S}_\lambda V \boldsymbol{S}_\lambda^T \tag{6.18}$$

Se concluye entonces que el estimador f_λ es sesgado, como lo son en general los estimadores en regresión no paramétrica y que tanto el estimador como su sesgo y su varianza dependen de λ. Así que la elección del parámetro de suavización es crucial, aunque en modo alguno trivial. Para encontrar el λ óptimo, los autores sugieren utilizar una estimación de la función de autocorrelación a partir de los residuales e_{ju}. Sin embargo, Rice & Silverman (1991) y Wang (2003) sugieren que los métodos tradicionales de validación cruzada que se basan en la idea de "dejar-una-observación-por-fuera" no son tan adecuados en situaciones como esta y que podría resultar más adecuado "dejar-un-individuo-por-fuera". Mayores detalles sobre la selección de λ en esta propuesta pueden consultarse en Bowman & Azzalini (1997, Pág. 139) y Diggle et al. (2002, Pág. 322).

Finalmente, Bowman & Azzalini (1997) no se detienen a estudiar cuidadosamente en el problema de la estimación de la varianza σ^2, a pesar de su gran importancia para efectos de inferencia, por ejemplo en la construcción de bandas de variabilidad. De hecho, los autores proponen estimar σ^2 utilizando la Expresión (6.19) para $k = 0$.

$$\hat{\gamma}_k = \frac{1}{n}\sum_{j=1}^{m}\sum_{u=1}^{n_j} e_{j,u}e_{j,u-k}, \quad k = 1, 2, \ldots, m \tag{6.19}$$

Pero la Expresión (6.19) para $k = 0$ no es más que el promedio de la suma de cuadrados de los residuales, estimador muy criticado porque no considera que los residuales dependen de λ. Una posible adaptación ha sido propuesta en algunos trabajos exploratorios (Pereira, Paz & Olaya 2007), acudiendo a los estimadores de Rice (1984) y Gasser et al. (1986) de tal manera que se estima la varianza siguiendo las

respuestas dentro de cada día, como si estuvieran dispuestas en el orden cronológico en el que son generadas, tal como se ven a la izquierda en la Figura 6.1.

Para efectos ilustrativos hemos estimado la función de regresión con la función sm.rm del paquete *sm* de **R** (Bowman & Azzalini 2010). Los resultados se ilustran en la Figura 6.4. Nótese que ni la nube de

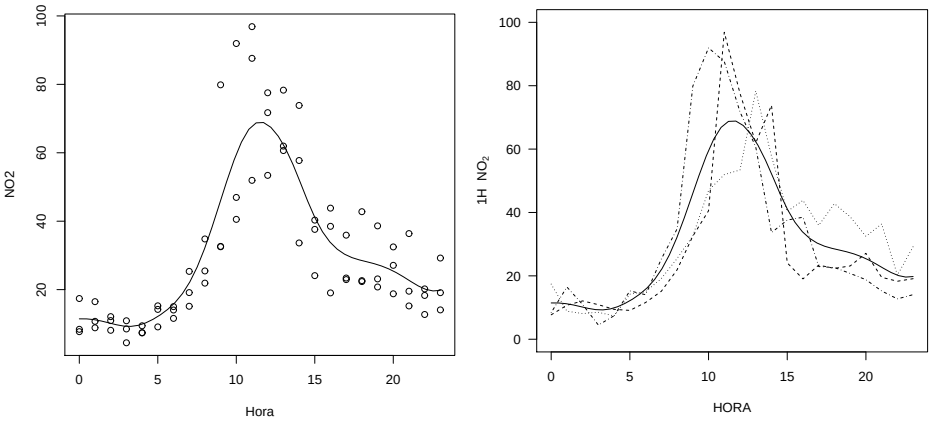


Figura 6.4: Curva suave ajustada a los datos de la contaminación 1H de NO para los días miércoles 20, jueves 21 y viernes 22 de enero de 2004 en el centro de Cali, usando el estimador de Bowman-Azzalini. Arriba se representa f sobre la nube de puntos y abajo sobre los perfiles

puntos ni la de perfiles apoyan el supuesto de igualdad de varianzas.

6.3. La aproximación de Eubank

Eubank (1999, Pág. 238) propone una solución diferente, más cercana a la idea de considerar los datos como respuestas múltiples que como medidas repetidas. Su idea se basa en el uso de splines.

Wahba (1990, pág. viii) y Green & Silverman (2000, pág. 14) describen un spline mecánico como una pieza metálica, plástica, de madera o de cualquier otro material flexible, que se ajusta a curvas adaptándose a su forma y que permite dibujar curvas *suaves*. Según estos autores, este tipo de herramienta se utilizó en el pasado para delinear cascos de barcos y para planear curvas de carrileras. Así que si fuera posible tener un objeto matemático que actuara como un spline mecánico que tuviera además adecuadas propiedades estadísticas, entonces podría utilizarse para ajustar curvas como las que nos proponemos en esta sección.

La versión más sencilla de un objeto matemático que se comporte

como un spline mecánico es llamado un *spline cúbico*. Supongamos que tenemos un conjunto de números reales $x_1, \ldots, x_n$ en un intervalo $[a, b]$, tales que $a < x_1 < x_2 < \ldots < x_n < b$. Una función s definida en $[a, b]$ es un spline cúbico si cumple las siguientes dos condiciones:

1. s es una cúbica en cada uno de los intervalos $(a, x_1), (x_1, x_2), \ldots, (x_n, b)$

2. La cúbicas se unen en los puntos x_i de tal manera que s y sus dos primeras derivadas con continuas en cada x_i y por lo tanto en todo el intervalo $[a, b]$

A los puntos x_i los llamaremos *nodos*.

Un spline cúbico en $[a, b]$ se llama un *spline cúbico natural* si se satisface que las dos primeras derivadas de s son iguales a cero en los puntos a y b. Estas condiciones las llamaremos condiciones de acotamiento natural e implican que s es lineal en los dos intervalos extremos (a, x_1) y (x_n, b).

Por otra parte, es posible demostrar (Green & Silverman 2000) que dados unos valores y_i, para un conjunto dado de puntos $x_1 < x_2 < \ldots < x_n$ en $[a, b]$, existe un único spline cúbico natural que satisface que $s(x_i) = y_i, \quad i = 1, \ldots, n$.

Supongamos que deseamos utilizar un spline cúbico natural para estimar f en nuestro Modelo (6.1). Para lograrlo, seguiremos a Eubank (1999) y Green & Silverman (2000), quienes proponen la cantidad $\int_0^1 f''(x)^2 dx$ como una medida natural de suavidad asociada con una función $f \in W_2^2[0, 1]$; al mismo tiempo, una medida de bondad de ajuste de los datos al modelo es la suma de cuadrados del error $n^{-1} \sum_{i=1}^n (y_i - f(x_i))^2$. Esto implica que una medida de la calidad de un estimador de f podría basarse en la suma convexa:

$$(1 - q)n^{-1} \sum_{i=1}^{n} (y_i - f(x_i))^2 + q \int_0^1 f''(x)^2 dx$$

con $0 < q < 1$.

Si hacemos $\lambda = q/(1 - q)$, la elección del estimador de f es equivalente a elegir f_λ que minimice la suma:

$$n^{-1} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_0^1 f''(x)^2 dx \tag{6.20}$$

sobre todas las funciones $f \in W_2^2[0,1]$. A este estimador f_λ lo llamaremos un estimador spline de f.

De la Expresión (6.20) se sigue que si λ es muy grande, entonces la estimación de la función de regresión será super-suavizada; lo contrario ocurre con un λ muy pequeño, que conduce a un estimador que interpola los datos.

Eubank (1999) encuentra que la solución a este problema de optimización es única y corresponde al estimador

$$f_\lambda = \sum_{j=1}^{n} \beta_{\lambda i} f_j \tag{6.21}$$

donde $\boldsymbol{\beta}_\lambda = (\beta_{\lambda 1}, \beta_{\lambda 2}, \ldots, \beta_{\lambda n})^T$ es la única solución con respecto a $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)^T$ del sistema de ecuaciones

$$(\boldsymbol{X}^T\boldsymbol{X} + n\lambda\boldsymbol{\Omega})\boldsymbol{c} = \boldsymbol{X}^T\boldsymbol{y} \tag{6.22}$$

donde $\boldsymbol{X} = \{f_j(x_i)\}_{i,j=1,2,\ldots,n}$, con $\{f_j\}_{j=1,2,\ldots}$ una colección de funciones que forman una base de $W_2^2[0,1]$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ y $\boldsymbol{\Omega} = \{\int_0^1 f_i''(x) f_j''(x) dx\}_{i,j=1,2,\ldots,n}$.

Las funciones $\{f_j\}_{j=1,2,\ldots,n}$ forman una base del conjunto de splines naturales. Sezer (2009) sugiere el uso de la siguiente base de splines cúbicos naturales:

$$\begin{aligned} f_1(x) &= 1 \\ f_2(x) &= x \\ f_{j+2}(x) &= d_j(x) - d_{n-1}(x), j = 1, 2, \ldots, n-2 \end{aligned} \tag{6.23}$$

donde:

$$d_j(x) = \frac{(x - x_j)_+^3 - (x - x_n)_+^3}{x_j - x_n}$$

y la función $(z)_+^3$ es la función truncada:

$$(z)_+^3 = \begin{cases} z^3, & \text{si } z \geq 0 \\ 0, & \text{si } z < 0 \end{cases}$$

En consecuencia, el vector de valores estimados es

$$\boldsymbol{f}_\lambda = (f_\lambda(x_1), f_\lambda(x_2), \ldots, f_\lambda(x_n))^T = \boldsymbol{S}_\lambda \boldsymbol{y},$$

donde tenemos que

$$\boldsymbol{S}_\lambda = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X}^T \tag{6.24}$$

Al estimador f_λ de f definido en la Ecuación (6.21) lo llamaremos un *estimador spline*. La elección del parámetro de suavización λ se hace usualmente con el estimador de validación cruzada generalizada GCV, usando la matriz $\boldsymbol{S}_\lambda$ definida en la Ecuación (6.24).

La solución $\boldsymbol{f_\lambda} = \boldsymbol{S}_\lambda\boldsymbol{y}$ basada en la matriz "hat" definida en la Ecuación (6.24) presume que tenemos una única respuesta y_i para cada x_i y todos los supuestos del Modelo (6.1), en particular igualdad de varianzas.

Una solución posible para manejar una situación como la que se describe en el Modelo (6.1), pero en presencia de heterocedasticidad, podría ser hallar f_λ que minimice la suma

$$n^{-1}\sum_{i=1}^{n} w_i(y_i - f(x_i))^2 + \lambda\int_0^1 f''(x)^2dx \tag{6.25}$$

con pesos positivos $w_i > 0, \quad i = 1, 2, \ldots, n$. Si usamos

$$w_i = [\mathrm{var}(y_i)]^{-1}, \quad i = 1, 2, \ldots, n$$

el estimador f_λ que minimice la suma (6.25) sería adecuado para el caso de un modelo heterocedástico.

Para el caso de un modelo homocedástico con respuestas múltiples, como sería el Modelo (6.4), se debería minimizar la suma

$$n^{-1}\sum_{j=1}^{m} n_j(\bar{y}_j - f(x_j))^2 + \lambda\int_0^1 f''(x)^2dx \tag{6.26}$$

con $n = \sum_{j=1}^{m} n_j$.

Si el Modelo (6.4) es heterocedástico, los ponderadores w_j pueden asociarse con la varianza muestral s_j^2 definida como

$$(n_j - 1)^{-1}\sum_{u=1}^{n_j}(y_{ju} - \bar{y}_j)^2 \tag{6.27}$$

para $j = 1, 2, \ldots, m$.

El estimador f_λ resultará de minimizar la suma

$$m^{-1}\sum_{j=1}^{m}\sum_{u=1}^{n_j} s_j^2(y_{ju} - f(x_j))^2 + \lambda\int_0^1 f''(x)^2dx \tag{6.28}$$

o la suma

$$m^{-1}\sum_{j=1}^{m}(\bar{y}_j - f(x_j))^2 + \lambda\int_0^1 f''(x)^2 dx \tag{6.29}$$

con $w_j = n_j/s_j^2, j = 1, 2, \ldots, m$.

En este caso la matriz $\boldsymbol{S}_\lambda$ será

$$\boldsymbol{S}_\lambda = \boldsymbol{J}\boldsymbol{A}_\lambda(\boldsymbol{J}^T\boldsymbol{J})^{-1}\boldsymbol{J}^T \tag{6.30}$$

donde $\boldsymbol{J}$ es una matriz $m \times n$ que es diagonal por bloques, con n_j unos en la diagonal del bloque j; $\boldsymbol{A}_\lambda = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + m\lambda\boldsymbol{\Omega})^{-1}\boldsymbol{J}^T\boldsymbol{W}$; y $\boldsymbol{W}$ es la matriz diagonal de pesos $w_j = n_j/s_j^2$.

La matriz $\boldsymbol{\Omega}$ se define como

$$\boldsymbol{\Omega} = \left\{\int_0^1 f_j''(x)f_u''(x)dx\right\}_{j,u=1,2,\ldots,m} \tag{6.31}$$

donde $\{f_j\}_{j=1,2,\ldots,m}$ es una base de splines cúbicos naturales, por ejemplo la base (6.23) propuesta por Sezer (2009).

Eubank (1999) no considera en su propuesta elementos tales como la estructura de dependencia de los errores, por lo que se asume que las mediciones sobre los individuos son independientes, así como las mediciones dentro de individuos. En nuestro Ejemplo 6.1, esto implicaría que las mediciones de un día a otro son independientes y que las mediciones de una hora a otra dentro de cada día son también independientes.

La Figura 6.5 ilustra la estimación de f basada en este procedimiento propuesto por Eubank, cuyo aporte más fuerte está en el uso de la medición de la varianza muestral de la Ecuación (6.27), que evidentemente permite aproximar la estimación de la varianza del modelo. Los supuestos de independencia deberían estudiarse más detenidamente.

6.4. ¿Qué hacer?

Sobre las propuestas de Bowman & Azzalini y Eubank

Las dos aproximaciones a la solución tienen dificultades que no han sido exploradas en la literatura. En el caso de la propuesta de Bowman & Azzalini (1997), la estimación de la varianza y de la estructura de dependencia de los errores es una fuente de trabajo. Existen algunos

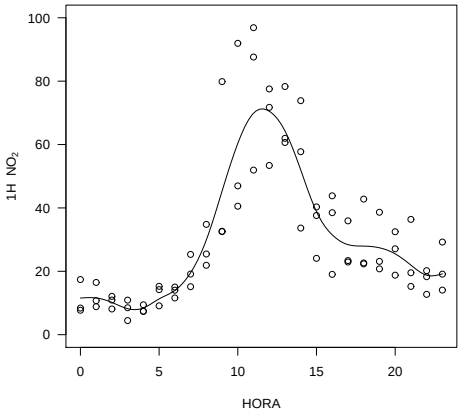


Figura 6.5: Estimación del comportamiento de la contaminación 1H de NO para los días miércoles 20, jueves 21 y viernes 22 de enero de 2004 en el centro de Cali, basada en la propuesta de Eubank

avances en el problema de la estimación de la varianza, pero hay muy poco trabajo en el tema de la estimación de las correlaciones y por lo tanto de la matriz V. Esta solución se asimila mucho al análisis de datos longitudinales, pero las sugerencias de Rice & Silverman (1991) que detalla Diggle et al. (2002) no parecen mejorar esta perspectiva. Además, se debería sustentar muy bien si en efecto tenemos o no datos longitudinales en el problema que nos ocupa.

En el caso de la propuesta de Eubank (1999) también se enfrenta el problema de asumir que las medidas dentro de cada individuo son independientes en nuestro caso. Desde luego hay casos en los que la metodología de Eubank es perfectamente adaptable, como es el caso de múltiples respuestas sobre individuos diferentes. Por ejemplo, si deseamos estudiar la relación entre la talla y el peso de hombres adultos, es evidente que dos hombres adultos de igual estatura pueden tener pesos diferentes. Y su estatura no cambiará significativamente con el tiempo, durante un largo periodo. En este caso tendremos respuestas múltiples de pesos de hombres adultos, para cada estatura. Así que es posible defender que no habrá dependencia "entre individuos" ni entre los errores. Pero nuestro problema no parece tener estas características específicas.

Una alternativa

En el caso del Ejemplo 6.1, en lugar de pensar que tenemos tres respuestas por cada punto de diseño, podríamos pensar que las

observaciones de cada día provienen en realidad de datos funcionales que hemos observado en puntos discretos. A los datos funcionales podríamos denotarlos $NO2_i(x), i = 1, 2, 3$. De esta manera, $NO2_1(x)$ sería el dato funcional correspondiente a la contaminación 1H de NO_2 el día miércoles 20 de enero de 2004. Esta idea se asemeja a la de los perfiles de la Figura 6.3, en la que hemos unido las observaciones con líneas rectas. De hecho, los perfiles podrían ser considerados datos funcionales en sí mismos, con la limitación de su falta de suavidad ya que su primera derivada será cero en varios puntos.

Una alternativa analítica para este problema sería usar las ideas de Ramsay & Silverman (2005), Ferraty & Vieu (2006) y Ramsay, Graves & Hooker (2010), entre otros, para ajustar a los datos un modelo de regresión funcional. Es bien sabido que los estimadores kernel, como los que proponen usar Bowman & Azzalini (1997), y los estimadores spline, que propone usar Eubank (1999), pueden verse como adecuaciones de modelos ajustados con series de Fourier. De hecho, el nombre kernel parece deberse al kernel Dirichlet del análisis de series de Fourier. Este es uno de los objetivos de trabajo futuro con las bases de datos de contaminación ambiental.

De acuerdo con Ramsay & Silverman (2005), los objetivos del análisis de datos funcionales son compartidos con otras ramas de la Estadística, entre otros los siguientes:

- ♢ Representar los datos para facilitar análisis posteriores
- ♢ Desplegar los datos para destacar varias de sus características
- ♢ Estudiar fuentes importantes de patrones de comportamiento y variación entre los datos
- ♢ Explicar la variación de una variable de respuesta a partir de información de una variable independiente

Todos estos objetivos son válidos para nuestro problema de la contaminación, por lo que ciertamente este tipo de análisis luce adecuado para nuestro problema.

Una primera etapa del análisis de datos funcionales es convertir los datos observados en una función g_i de tal manera que sea posible calcular $g(x)$ para cualquier x, lo que se haría por interpolación si las respuestas fueran medidas sin error, o, más comúnmente, por suavización. La Figura 6.6 muestra los datos funcionales para este problema, obtenidos usando suavización spline.

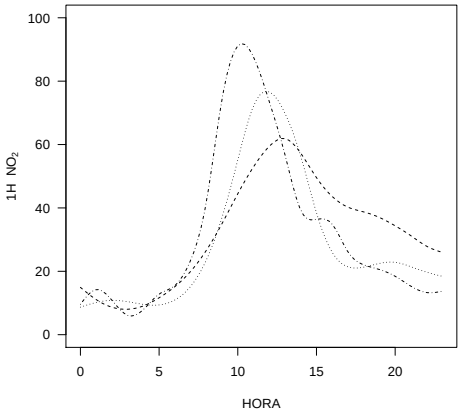


Figura 6.6: Tres datos funcionales de la contaminación 1H de NO para los días miércoles 20, jueves 21 y viernes 22 de enero de 2004 en el centro de Cali

En realidad los datos disponibles para el análisis de la contaminación por NO_2 en Cali son varias decenas, por lo que la Figura 6.6 es solo una primera aproximación gráfica a un problema mucho más complejo. Es decir, el análisis de esta información usando las técnicas del análisis funcional es aún un problema en curso.

BIBLIOGRAFÍA

Arango, J. & Calderón, G. (2006), *Notas de clase: Ecuaciones Diferenciales*, Departamento de Matemáticas, Universidad del Valle, Cali, Colombia.

Barrientos, A. F., Olaya, J. & Gonzalez, V. M. (2007), 'Un modelo spline para el pronóstico de la demanda de energía eléctrica', *Revista Colombiana de Estadística* **30**(2), 187–202.

Bartle, R. G. (1995), *The Elements of Integration and Lebesgue Measure*, Wiley, New York, NY.

Benedetti, G. (1975), 'Kernel estimation of regresion functions', *Proceedings in Computer Science and Statistics: 8^{th} annual symposium on the interface* pp. 405–412.

Bowman, A. W. & Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis. The Kernel Approach with S-plus Illustrations*, Oxford.

Bowman, A. W. & Azzalini, A. (2010), *R package* `sm`*: nonparametric smoothing methods (version 2.2-4)*, University of Glasgow, UK and Università di Padova, Italia.

Breiman, L. (1991), 'The π method for estimating multivariate functions from noisy data', *Technometrics* **33**(2), 125–160.

Brown, L. D. & Levine, M. (2007), 'Variance estimation in nonparametric regression via the difference sequence method', *The Annals of Statistics* **35**, 2219–2232.

Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association* **74**, 829–836.

Delicado, P. (2008), *Curso de Modelos no Paramétricos*, UPC.

Diggle, P. J., Heagerty, P., Liang, K.-Y. & Zeger, S. L. (2002), *Analysis of Longitudinal Data*, 2nd. edn, Oxford.

Draper, N. R. & Smith, H. (1966), *Applied Regression Analysis*, John Wiley & Sons, New York, NY.

Draper, N. R. & Smith, H. (1998), *Applied Regression Analysis*, 3ra. edn, John Wiley & Sons, New York, NY.

Epanechnikov, V. (1969), 'Non-parametric estimation of a multivariate probability density', *Theory of Probability and its Applications* **14**, 153–158.

Eubank, R. L. (1999), *Nonparametric Regression and Spline Smoothing*, second edn, Marcel Dekker, New York, NY.

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

Ferraty, F., Nuñez Antón, V. & Vieu, P. (2001), *Regresión no Paramétrica: Desde la Dimensión Uno Hasta la Dimensión Infinita*, UPV/EHU.

Ferraty, F. & Vieu, P. (2006), *Nonparametric Functional Data Analysis Theory and Practice*, Springer.

Friedman, J. H. & Stuetzle, W. (1981), 'Projection pursuit regression', *Journal of the American Statistical Association, Theory and Methods* **76**(376), 817–823.

Fromkorth, A. & Kohler, M. (2011), 'Estimating the mean and covariance structure nonparametrically when the data are curves', *Journal of the Royal Statistical Society, Series B* **141**, 172–188.

Gasser, T. & Müller, H. G. (1979), *Smoothing Techniques for Curve Estimation*, Springer, Heidelberg.

Gasser, T., Sroka, L. & Jennen-Steinmetz, C. (1986), 'Residual variance and residual pattern in nonlinear regression', *Biometrika* **73**(3), 625–633.

Gradshtein, I. S. & Ryzhik, I. M. (1980), *Tables of Integrals, Series, and Products*, Academic Press.

Graybill, F. A. (1976), *Theory and Application of the Linear Model*, Wadsworth & Brooks, Pacific Grove, California.

Graybill, F. A. (2001), *Matrices with Applications in Statistics*, second edn, Cengage Learning.

Green, P. J. & Silverman, B. W. (2000), *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach*, Chapman & Hall/CRC, Boca Raton, FL.

Hall, P. (1989), 'On projection pursuit regression', *The Annals of Statistics* **17**(2), 573–588.

Hall, P., Kay, J. W. & Titterington, D. M. (1990), 'Asymptotically optimal difference-based estimation of variance in nonparametric regression', *Biometrika* **77**(3), 521–528.

Härdle, W. (1990*a*), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, UK.

Härdle, W. (1990*b*), *Smoothing Techniques, with Applications in S*, Cambridge University Press, Cambridge, UK.

Härdle, W., Hall, P. & Ichimura, H. (1993), 'Optimal smoothing in single-index models', *The Annals of Statistics* **21**(1), 157–178.

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman & Hall-CRC, Boca Raton, FL.

Junger, W. & de Leon, A. P. (2011), *ares: Environment air pollution epidemiology: a library for time series analysis*. R package version 0.7.2.
URL: `http://CRAN.R-project.org/package=ares`

Kohler, M. & Krzyúzak, A. (2010), 'Pricing of american options in discrete time using least squares estimates with complexity penalties'.

Kreyszig, E. (1989), *Introductory Functional Analysis with Applications*, Wiley, New York, NY.

Levine, M. (2006), 'Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: A possible approach', *Comput. Stat. Data Anal.* **50**, 3405–3431.

Li, K.-C. (1991), 'Sliced inverse regression for dimension reduction', *Journal of the American Statistical Association, Theory and Methods* **86**(414), 316–327.

Müller, H. G. (1988), *Nonparametric Regression Analysis of Longitudinal Data*, Springer, New York, NY.

Nadaraya, E. A. A. & Seckler, B. T. (1964), 'On estimating regression', *Theory of Probability and its Applications (Transl of Teorija Verojatnostei i ee Primenenija)* **9**, 141–142.

Nelder, J. A. & McCoullagh, P. (1989), *Generalized linear models*, 2 edn, Chapman & Hall/CRC.

Nelder, J. A. & Wedderburn, R. W. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society Series A* **135**(3), 370–384.

Neter, J., Wasserman, W. & Kutner, M. H. (1990), *Applied Linear Statistical Models*, 3ra. edn, Irwin, Burr Ridge, Illinois.

Pereira, L. A., Paz, M. C. & Olaya, J. (2007), Estimación de la varianza en regresión no-paramétrica: El efecto de poseer múltiples observaciones por punto de diseño, *in* '17mo. Simposio de Estadística', Universidad Nacional de Colombia.

Priestley, M. & Chao, M. (1972), 'Non-parametric function fitting', *Journal of the Royal Statistical Society, Series B* **34**, 385–392.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: `http://www.R-project.org/`

Ramsay, J. O., Graves, S. & Hooker, G. (2010), *Functional Data Analysis with **R** and **MATLAB***, Springer.

Ramsay, J. O. & Silverman, B. W. (2005), *Functional Data Analysis*, 2nd. edn, Springer.

Rice, J. A. (1984), 'Bandwidth choice for nonparametric regression', *The Annals of Statistics* **12**(4), 1215–1230.

Rice, J. A. & Silverman, B. W. (1991), 'Estimating the mean and covariance structure nonparametrically when the data are curves', *Journal of the Royal Statistical Society, Series B* **53**, 233–243.

Searle, S. R. (1971), *Linear Models*, John Wiley & Sons, New York, NY.

Seber, G. A. F. (1977), *Linear Regression Analysis*, John Wiley & Sons, New York, NY.

Sezer, A. (2009), 'Assesing the quality of the natural cubic spline approximation', *Proceedings of the 8th WSEAS International Conference on SYSTEM SCIENCE and SIMULATION in ENGINEERING* pp. 186–190.

Simonoff, J. F. (1996), *Smoothing Methods in Statistics*, Springer, New York, NY.

Singh, S. (1997), *El enigma de Fermat*, Planeta.

Sockett, E. B., Daneman, D., Clarson, C. & Ehrich, R. M. (1987), 'Factors affecting and patterns of residual insulin secretion daring the first year of type i (insulin dependent) diabetes mellitus in children', *Diabet.* **30**, 453–459.

Takezawa, K. (2006), *Introduction to Nonparametric Regression*, Wiley.

Tong, T. & Wang, Y. (2005), 'Estimating residual variance in nonparametric regression using least squares', *Biometrika* **93**, 821–830.

Wahba, G. (1990), *Spline Models for Observational data*, CBMS-NSF Series, SIAM.

Wang, J.-L. (2003), 'Nonparametric regression analysis of longitudinal data'.
URL: `http://www.stat.ucdavis.edu/~wang/paper/EOB3.pdf`

Watson, G. (1964), 'Smooth regression analysis', *Sankhya, series A* **26**, 359–372.

ÍNDICE ALFABÉTICO

Universidad del Valle

Programa Editorial

Ciudad Universitaria, Meléndez
Cali, Colombia
Teléfonos: (+57) 2 321 2227
321 2100 ext. 7687
http://programaeditorial.univalle.edu.co
programa.editorial@correounivalle.edu.co

¡Síguenos!

programaeditorialunivalle