Curso de Modelos no Paramétricos

Pedro Delicado

Departament d'Estadística i Investigació Operativa

Universitat Politècnica de Catalunya

14 de septiembre de 2008

Índice general

Prefacio **v**

1. Contrastes no paramétricos clásicos **1**
1.1. Introducción 1
1.2. Contrastes de bondad de ajuste 1
1.2.1. La función de distribución empírica 2
1.2.2. El contraste de Kolmogorov-Smirnov 7
1.2.3. Bondad de ajuste a un modelo paramétrico. 9
1.3. Contrastes de localización 9
1.3.1. El test del signo 10
1.3.2. Test de Wilcoxon de los rangos signados 12
1.4. Comparación de dos muestras independientes 15
1.4.1. Test de Kolmogorov-Smirnov para dos muestras 16
1.4.2. Test de Mann-Whitney-Wilcoxon 17
1.5. Comparación de más de dos muestras 19
1.5.1. Muestras independientes: Test de Kruskal-Wallis 19
1.5.2. Muestras relacionadas: Test de Friedman 20
1.6. Medida de la dependencia 21
1.6.1. Coeficiente τ de Kendall 22
1.6.2. Coeficiente de correlación de rangos de Spearman . . . 23
1.7. Comentarios finales 25

2. Introducción a los métodos de suavizado **29**
2.1. Introducción 29
2.2. Usos de los métodos de suavizado. 35

3. Estimación no paramétrica de la densidad **39**
3.1. La estimación de la densidad 39
3.2. El histograma y el polígono de frecuencias 40
3.2.1. Motivación del histograma 41

3.2.2. Características del histograma 42
3.2.3. Propiedades locales del estimador histograma 44
3.2.4. Propiedades globales del estimador histograma 47
3.2.5. Elección del parámetro de suavizado b 49
3.2.6. El polígono de frecuencias 51
3.2.7. Comportamiento asintótico del polígono de frecuencias 52
3.3. Estimador núcleo de la densidad 53
3.3.1. Comportamiento asintótico del estimador núcleo de la densidad 59
3.3.2. Problemas de los estimadores núcleo y algunas soluciones 68
3.4. Selección automática del parámetro de suavizado 77
3.4.1. Regla de referencia a la normal 77
3.4.2. Sobresuavizado 77
3.4.3. Validación cruzada por mínimos cuadrados 79
3.4.4. Plug-in directo 81
3.4.5. Validación cruzada por máxima verosimilitud 83
3.4.6. Otros métodos 83
3.5. Estimación de la densidad multivariante 85
3.5.1. Elección de la matriz ventana 89
3.5.2. Representación de densidades tri-variantes 91
3.5.3. La maldición de la dimensionalidad 93
3.6. Inferencia basada en la estimación de la densidad 95
3.6.1. Bandas de variabilidad 95
3.6.2. Contraste de normalidad 97
3.6.3. Bandas de referencia normal 100
3.6.4. Contraste de independencia 101
3.6.5. Bootstrap en la estimación de la densidad 102
3.6.6. Contraste de igualdad de distribuciones 103
3.6.7. Discriminación no paramétrica basada en estimación de la densidad 104
3.7. Otros estimadores de la densidad 108
3.7.1. Los k vecinos más cercanos 108
3.7.2. Desarrollos en series de funciones ortogonales 110
3.7.3. Máxima verosimilitud penalizada 111
3.7.4. Verosimilitud local 113
3.7.5. Representación general 113
3.8. Software 113
3.8.1. Estimación de la densidad en R 113
3.8.2. Estimación de la densidad en MATLAB 113

4. Estimación de la función de regresión 115
4.1. El modelo de regresión no paramétrica 117
4.2. Estimadores núcleo y polinomios locales 119
4.2.1. Derivación directa del estimador núcleo de la regresión 125
4.2.2. Expresión matricial del estimador por polinomios locales 127
4.2.3. Propiedades locales de los estimadores por polinomios locales 128
4.2.4. Comportamiento en la frontera del soporte de x 131
4.2.5. Elección del grado del polinomio local 131
4.3. Elección del parámetro de suavizado 134
4.3.1. Error de predicción en una muestra test 135
4.3.2. Validación cruzada 135
4.3.3. Validación cruzada generalizada. 136
4.3.4. Plug-in 139
4.3.5. Comportamiento asintótico de selectores de h 143
4.3.6. Ventana variable 143
4.4. Verosimilitud local 144
4.4.1. Discriminación no paramétrica mediante regresión binaria local 144
4.4.2. Modelo de verosimilitud local 148
4.5. Inferencia en el modelo de regresión no paramétrica 151
4.5.1. Bandas de variabilidad 152
4.5.2. Contraste de ausencia de efectos 153
4.5.3. Contraste de un modelo lineal 155
4.5.4. Contraste de un modelo lineal generalizado 156
4.5.5. Igualdad de curvas de regresión 157

5. Estimación por splines 161
5.1. Estimación mínimo cuadrática penalizada 161
5.2. Splines y splines cúbicos. Interpolación por splines 163
5.3. Suavizado por splines 166
5.4. Propiedades del estimador spline de $m(x)$ 169
5.5. B-splines 170
5.6. Ajuste de un modelo no paramétrico general 173

6. Regresión múltiple y modelo aditivo generalizado 175
6.1. Regresión múltiple 175
6.2. Modelos aditivos 179
6.3. Regresión *projection pursuit* 182
6.4. Modelos aditivos generalizados 183
6.5. Modelos semiparamétricos 186

A. Apéndice. Algunos conceptos y resultados matemáticos 187

Referencias 190

Prefacio

Modelos paramétricos versus no paramétricos

Sea X variable aleatoria con distribución de probabilidad dada por la función de distribución F. Diremos que la v.a. X sigue un MODELO PARAMÉTRICO si su distribución de probabilidad F pertenece a una familia de distribuciones indexada por un parámetro θ de dimensión finita:

$$X \sim F \in \mathcal{F}_\Theta = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}.$$

La familia de distribuciones $\mathcal{F}_\Theta$ recibe el nombre de MODELO ESTADÍSTICO PARAMÉTRICO.

Diremos que la v.a. X sigue un MODELO ESTADÍSTICO NO PARAMÉTRICO si sobre su distribución F únicamente se suponen algunas condiciones de regularidad. Algunos ejemplos de estas condiciones son los siguientes:

- F es una función de distribución absolutamente continua,
- F es simétrica en torno a su mediana,
- F tiene función de densidad f con dos derivadas continuas.

Las restricciones impuestas sobre F indican que esta distribución pertenece a un subconjunto de todas las posibles distribuciones de probabilidad, pero este subconjunto tiene dimensión infinita (no se puede indexar por un parámetro de dimensión finita).

Métodos no paramétricos

Son métodos de inferencia estadística válidos cuando no se hacen hipótesis paramétricas sobre la distribución de los datos. Distinguiremos dos familias de métodos. La primera fue desarrollada principalmente en las décadas de los 40 y 50 del siglo XX, y la segunda en el último tercio de ese siglo.

Métodos no paramétricos clásicos

Tienen por objetivo hacer inferencia sobre la distribución de probabilidad F de X o sobre alguna característica suya que esté bien definida sea cual sea la distribución F (por ejemplo, la mediana o el rango intercuartílico de F).

Como no se conoce la distribución F los métodos que se proponen se basan en estadísticos cuya distribución en el muestreo no depende de F. Por ello se conocen como MÉTODOS LIBRES DE LA DISTRIBUCIÓN DE LOS DATOS, o MÉTODOS DE DISTRIBUCIÓN LIBRE (una mala traducción del término *distribution-free* en inglés).

Éstos son los métodos que trataremos en el Capítulo 1. Concretamente nos centraremos en contrastes de hipótesis no paramétricos.

Estimación no paramétrica de curvas

Son técnicas que permiten estimar funciones relacionadas con la distribución de probabilidad de los datos. Por ejemplo se puede tener interés en estimar la función de distribución $F(x)$, la función de densidad $f(x)$, la tasa de fallo $\lambda(x) = f(x)/(1-F(x))$, la función de regresión $m(x) = E(Y|X = x)$ o la varianza condicional $\sigma^2(x) = V(Y|X = x)$. A estas técnicas se dedicarán los restantes capítulos.

Capítulo 1

Contrastes no paramétricos clásicos

REFERENCIAS: Pratt y Gibbons (1981), Gibbons y Chakraborti (1992), Gibbons (1993a), Gibbons (1993b), Gibbons (1997), Hollander y Wolfe (1999), Leach (1982)

1.1. Introducción

En este capítulo presentamos algunos de los contrastes de hipótesis no paramétricos clásicos. Todos tienen en común que no precisan hacer hipótesis paramétricas sobre la distribución de probabilidad F de los datos, pues se basan en estadísticos cuya distribución en el muestreo no depende de F. Son por tanto CONTRASTES LIBRES DE LA DISTRIBUCIÓN DE LOS DATOS (*distrbution-free tests*).

Veremos en primer lugar contrastes de bondad de ajuste basados en la distribución empírica de los datos. Después veremos contrastes de localización para una muestra (o para dos muestras apareadas), contrastes que igualdad de dos muestras, versiones no paramétricas de los contrastes ANOVA clásicos y, por último, medidas no paramétricas de la dependencia entre dos variables.

1.2. Contrastes de bondad de ajuste

Nos planteamos el problema de saber si una variable aleatoria sigue o no una distribución determinada. Sea X v.a. con función de distribución F desconocida. Sea F_0 una función de distribución conocida. Se desea contrastar

$$H_0 : F = F_0 \text{ frente a } H_1 : F \neq F_0.$$

Para ello se dispone de una muestra aleatoria simple (m.a.s.) $X_1, \ldots, X_n$ de X. También consideramos las hipótesis alternativas unilaterales $H_1 : F(x) > F_0(x)$ para todo x, o $H_1 : F(x) < F_0(x)$ para todo x.

Vamos a estudiar el contraste de Kolmogorov-Smirnov (existen otras formas de realizar contrastes de bondad de ajuste, por ejemplo los contrastes de la χ^2, basados en la categorización de los datos).

El contraste de Kolmogorov-Smirnov se basa en calcular una distancia entre la función de distribución empírica de los datos, F_n, y la función de distribución F_0 postulada bajo H_0. Recordemos la definición y propiedades de la función de distribución empírica.

1.2.1. La función de distribución empírica

Sea la variable aleatoria X con función de distribución F. Consideramos una muestra aleatoria simple de tamaño n de X, es decir, $X_1, \ldots, X_n$ v.a.i.i.d. con distribución dada por F. Sea $x_1, \ldots, x_n$ una realización de esa m.a.s.

Se llama FUNCIÓN DE DISTRIBUCIÓN EMPÍRICA a la función

$$F_n(x) = \frac{1}{n}\#\{x_i \leq x : i = 1 \ldots n\} = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(x_i),$$

donde

$$I_{(-\infty,x]}(x_i) = \begin{cases} 1, & \text{si } x_i \leq x \\ 0, & \text{si } x_i > x, \end{cases}$$

que a cada número real x le asigna la proporción de valores observados que son menores o iguales que x.

Es inmediato comprobar que la función F_n así definida es una función de distribución:

1. $F_n(x) \in [0, 1]$ para todo $x \in \mathbb{R}$.
2. F_n es continua por la derecha.
3. F_n es no decreciente.
4. $\lim_{x \longrightarrow -\infty} F_n(x) = 0$.
5. $\lim_{x \longrightarrow \infty} F_n(x) = 1$.

Concretamente, F_n es la función de distribución de una variable aleatoria discreta (que podemos llamar X_e) que pone masa $1/n$ en cada uno de los n puntos x_i observados:

x_i	x_1	x_2	$\cdots$	x_n
$p_i = \mathrm{P}(X_e = x_i)$	$1/n$	$1/n$	$\cdots$	$1/n$

A la distribución de X_e se le llama DISTRIBUCIÓN EMPÍRICA asociada al conjunto de valores $\{x_1, \ldots, x_n\}$.

Obsérvese que si fijamos el valor de x y dejamos variar la muestra, lo que obtenemos es una variable aleatoria. En efecto, se tiene entonces que

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i),$$

donde

$$I_{(-\infty,x]}(X_i) = \begin{cases} 1, & \text{si } X_i \leq x \\ 0, & \text{si } X_i > x \end{cases}$$

y, por lo tanto, cada término $I_{(-\infty,x]}(X_i)$ es una variable aleatoria de Bernoulli con probabilidad de éxito

$$p = \mathrm{P}(I_{(-\infty,x]}(X_i) = 1) = \mathrm{P}(X_i \leq x) = F(x).$$

De ahí se deduce que F_n es una variable aleatoria y que $nF_n(x)$ tiene distribución binomial con parámetros n y $p = F(x)$.

De lo anterior se sigue que la función de distribución empírica es un proceso estocástico: si consideramos un espacio probabilístico $(\Omega, \mathcal{A}, P)$ donde están definidas las sucesiones de variables aleatorias $\{X_n\}_{n\geq 1}$ a partir de las cuales definiremos la función de distribución empírica, tenemos que

$$\begin{array}{rccl} F_n: & (\Omega, \mathcal{A}, P) \times (\mathbb{R}, \mathcal{B}) & \longrightarrow & [0,1] \\ & (\omega, x) & \longrightarrow & F_n(x)(\omega) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i(\omega)). \end{array}$$

Fijado x, $F_n(x)(\cdot) : (\Omega, \mathcal{A}, P) \longrightarrow [0,1]$ es una variable aleatoria. Fijado ω, $F_n(\cdot)(\omega) : \mathbb{R} \longrightarrow [0,1]$ es una función de distribución (en la notación usual se omite la dependencia de $\omega \in \Omega$). Por lo tanto, la función de distribución empírica es una *función de distribución aleatoria*.

El siguiente teorema recoge algunas de las propiedades de la función de distribución empírica.

Teorema 1.1 *Sea $\{X_n\}_{n\geq 1}$, sucesión de variables aleatorias independientes e idénticamente distribuidas definidas en el espacio de probabilidad $(\Omega, \mathcal{A}, P)$ con función de distribución común F. Se denota por F_n la función de distribución empírica obtenida de las n primeras variables aleatorias $X_1, \ldots, X_n$. Sea $x \in \mathbb{R}$. Se verifica lo siguiente:*

(a) $P(F_n(x) = \frac{j}{n}) = \binom{n}{j} F(x)^j (1 - F(x))^{n-j}$, $j = 0, \ldots, n$.

(b) $E(F_n(x)) = F(x)$, $Var(F_n(x)) = (1/n)F(x)(1-F(x))$.

(c) $F_n(x) \longrightarrow F(x)$ *casi seguro.*

(d)

$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1-F(x))}} \longrightarrow_D Z,$$

donde Z es una variable aleatoria con distribución normal estándar y la convergencia es convergencia en distribución.

Demostración: Los apartados (a) y (b) son consecuencia inmediata del hecho de que $nF_n(x) \sim B(n, p = F(x))$. Por otro lado, si definimos $Y_i = I_{(-\infty,x]}(X_i)$, se tiene que $F_n(x) = \bar{Y}_n$, la media aritmética de las variables aleatorias $Y_1, \ldots, Y_n$. Así, el apartado (c) es una aplicación inmediata de la ley fuerte de los grandes números y el apartado (d) es consecuencia del teorema central de límite. □

El siguiente teorema refuerza el resultado (c) anterior, puesto que afirma que la convergencia de $F_n(x)$ a $F(x)$ se da uniformemente.

Teorema 1.2 (Teorema de Glivenko-Cantelli) *Sea $\{X_n\}_{n\geq 1}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas definidas en el espacio de probabilidad $(\Omega, \mathcal{A}, P)$ con función de distribución común F. Se denota por F_n la función de distribución empírica obtenida de las n primeras variables aleatorias $X_1, \ldots, X_n$. Entonces,*

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \longrightarrow 0 \text{ casi seguro.}$$

Demostración: Presentamos aquí la demostración que hacen Vélez y García (1993), p. 36. (otras demostraciones pueden encontrarse en García-Nogales 1998, p. 88, y en Cristóbal 1992, p. 66). En el Teorema 1.1 se probó que, por la ley fuerte de los grandes números, $F_n(x) \longrightarrow F(x)$ casi seguro, es decir, para cada $x \in \mathbb{R}$ existe $A_x \in \mathcal{A}$ tal que $\mathrm{P}(A_x) = 1$ y $\lim_n F_n(x)(\omega) = F(x)$ si $\omega \in A_x$. Se ha denotado por $F_n(x)(\omega)$ a la función de distribución empírica obtenida al observar $X_1(\omega), \ldots, X_n(\omega)$, siendo ω un elemento del espacio Ω. De la ley fuerte de los grandes números también se sigue (tomando ahora $I_{(-\infty,x)}$ en vez de $I_{(-\infty,x]}$) que para cada $x \in \mathbb{R}$, existe $B_x \in \mathcal{A}$ tal que $\mathrm{P}(B_x) = 1$ y $\lim_n F_n(x^-)(\omega) = F(x^-)$ si $\omega \in B_x$, donde $g(x^-)$ denota el límite por la izquierda de una función g en x.

Para cada número natural k, y cada $j = 1, \ldots, k$, se consideran los puntos

$$x_{jk} = \text{mín}\left\{x \in \mathbb{R} : F(x^-) \leq \frac{j}{k} \leq F(x)\right\}$$

y los sucesos de $\mathcal{A}$ siguientes:

$$A_{jk} = A_{x_{jk}} = \{w \in \Omega : F_n(x_{jk}) \longrightarrow F(x_{jk})\}$$

$$B_{jk} = B_{x_{jk}} = \{w \in \Omega : F_n(x_{jk}^-) \longrightarrow F(x_{jk}^-)\}$$

$$D_k = \bigcap_{j=1}^{k}(A_{jk} \cap B_{jk}),\ D = \bigcap_{k=1}^{\infty} D_k.$$

D_k es el suceso definido por la condición de que la función de distribución empírica converja a la teórica para todos los puntos x_{jk} (y también para los límites por la izquierda), para un k fijo. D es el suceso en que esto ocurre simultáneamente para todo k. Según la ley fuerte de los grandes números, $\mathrm{P}(A_{jk}) = \mathrm{P}(B_{jk}) = 1$ para todo j y todo k, luego $\mathrm{P}(D_k) = 1$ para todo k y, por tanto, $\mathrm{P}(D) = 1$.

Obsérvese que si $x \in [x_{jk}, x_{(j+1)k})$, por ser F y F_n funciones de distribución se tiene que

$$F(x_{jk}) \leq F(x) \leq F(x_{(j+1)k}^-),\ \text{y}\ F_n(x_{jk}) \leq F_n(x) \leq F_n(x_{(j+1)k}^-).$$

Como además $F(x_{(j+1)k}^-) - F(x_{jk}) \leq 1/k$,

$$F_n(x) - F(x) \leq F_n(x_{(j+1)k}^-) - F(x_{jk}) \leq F_n(x_{(j+1)k}^-) - F(x_{(j+1)k}^-) + \frac{1}{k}$$

y

$$F_n(x) - F(x) \geq F_n(x_{jk}) - F(x_{(j+1)k}^-) \geq F_n(x_{jk}) - F(x_{jk}) - \frac{1}{k}$$

con lo cual, si $\delta_n^{(k)}$ es la mayor entre todas las diferencias $|F_n(x_{jk}) - F(x_{jk})|$ y $|F_n(x_{jk}^-) - F(x_{jk}^-)|$ (para n y k fijos), se tiene que

$$F_n(x) - F(x) \leq \delta_n^{(k)} + \frac{1}{k}\ \text{y}\ F_n(x) - F(x) \geq -\delta_n^{(k)} - \frac{1}{k}$$

Así, para cualquier $k \in \mathbb{N}$,

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \leq \delta_n^{(k)} + \frac{1}{k}.$$

Obsérvese que si se verifica el suceso D, para cualquier $k \in \mathbb{N}$ y cualquier $\varepsilon > 0$, se tiene que $\delta_n^{(k)} < \varepsilon$ a partir de un cierto n, de forma que

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| < \varepsilon + \frac{1}{k}$$

a partir de cierto n. Por lo tanto,

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \longrightarrow_n 0$$

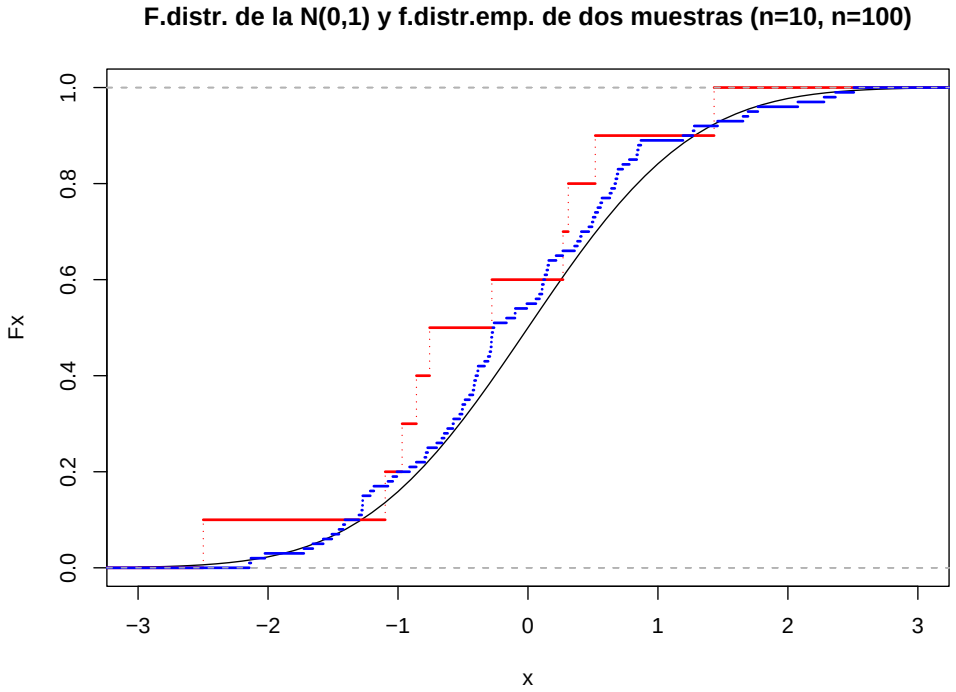
siempre que se verifique D. Como $\mathrm{P}(D) = 1$, se sigue que

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \longrightarrow_n 0 \text{ casi seguro.}$$

□

Ejemplo 1.1

En la figura siguiente se muestra la función de distribución de una variable aleatoria $N(0,1)$ y la función de distribución empírica de dos muestras de esa variable aleatoria una de tamaño $n = 10$ (la más alejada de la teórica) y la otra de tamaño $n = 100$. Se aprecia que cuando n crece la proximidad entre la función de distribución empírica y la teórica es cada vez mayor.



1.2.2. El contraste de Kolmogorov-Smirnov

El Teorema de Glivenko-Cantelli da pie a basar el contraste de bondad de ajuste en el estadístico

$$D_n = \sup_x |F_n(x) - F_0(x)|$$

que se llama ESTADÍSTICO BILATERAL DE KOLMOGOROV-SMIRNOV. También serán útiles los ESTADÍSTICOS UNILATERALES DE KOLMOGOROV-SMIRNOV:

$$D_n^+ = \sup_x (F_n(x) - F_0(x)),\ D_n^- = \sup_x (F_0(x) - F_n(x))$$

para los contrastes unilaterales. Bajo H_0 se tiene que estos estadísticos convergen a 0 casi seguro cuando $n \longrightarrow \infty$. Observar que

$$D_n = \text{máx}\{D_n^+, D_n^-\}.$$

La siguiente tabla resume la forma de llevar a cabo los contrastes de bondad de ajuste:

Hipótesis nula	Hipótesis alternativa	Región crítica (nivel α)	p-valor
$F(x) = F_0(x)$	$F(x) \neq F_0(x)$	$D_n^{Obs} > D_{n,\alpha}$	$P(D_n > D_n^{Obs})$
$F(x) = F_0(x)$	$F(x) > F_0(x)$	$D_n^{+,Obs} > D_{n,\alpha}^+$	$P(D_n^+ > D_n^{+,Obs})$
$F(x) = F_0(x)$	$F(x) < F_0(x)$	$D_n^{-,Obs} > D_{n,\alpha}^-$	$P(D_n^- > D_n^{-,Obs})$

Los valores D_n^{Obs}, $D_n^{+,Obs}$ y $D_n^{-,Obs}$ son los valores observados de los estadísticos D_n, D_n^+ y D_n^-, respectivamente. Los valores $D_{n,\alpha}$, $D_{n,\alpha}^+$ y $D_{n,\alpha}^-$ son los que dejan a su derecha una probabilidad α en las distribuciones bajo H_0 de D_n, D_n^+ y D_n^-, respectivamente.

Para encontrar los valores $D_{n,\alpha}$, $D_{n,\alpha}^+$ y $D_{n,\alpha}^-$ o calcular los p-valores es necesario conocer la distribución de los estadísticos D_n, D_n^+ y D_n^-. Vamos a estudiar estas distribuciones a continuación.

La siguiente proposición establece que si F_0 es absolutamente continua y estrictamente creciente los contrastes basados en estos estadísticos son de distribución libre.

Proposición 1.1 *Supongamos que F_0 es absolutamente continua y estrictamente creciente. Bajo H_0 la distribución de D_n, D_n^+ y D_n^- no depende de F_0.*

Demostración: Recordar que si F_0 es absolutamente continua y estrictamente creciente, se tienen las siguientes propiedades:

- Si $X \sim F_0$ entonces $F_0(X) \sim U([0,1])$.
- Si $U \sim U([0,1])$ entonces $F_0^{-1}(U) \sim F_0$.

Observar que la función de distribución empírica puede reescribirse así:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(F_0^{-1}(U_i)) =$$

$$\frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,F_0(x)]}(U_i) = F_n^U(F_0(x)),$$

donde $U_1,\ldots,U_n$ es una m.a.s. de una $U([0,1])$ y F_n^U es su función de distribución empírica. Así,

$$D_n = \sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)| = \sup_{x\in\mathbb{R}} |F_n^U(F_0(x)) - F_0(x)| = \sup_{u\in[0,1]} |F_n^U(u) - u|,$$

que es el valor del estadístico de Kolmogorov-Smirnov calculado a partir de una m.a.s. de una $U([0,1])$. Por lo tanto la distribución de D_n no depende de F_0. Análogos argumentos pueden hacerse para D_n^+ y D_n^-. □

Distribución exacta.
La distribución exacta de D_n, D_n^+ y D_n^- puede calcularse para cualquier tamaño muestral n utilizando técnicas estándar de cálculos de probabilidades a partir de la función de densidad conjunta de la variable aleatoria multivariante $(U_1,\ldots,U_n)$. También pueden aproximarse esas distribuciones mediante simulación. Estas distribuciones están tabuladas en muchos libros de estadística (ver Gibbons 1997 o Hollander y Wolfe 1999, por ejemplo).

Obsérvese que la distribución de D_n^- coincide con la de D_n^+ para cualquier tamaño muestral.

Distribución asintótica.
Si el tamaño muestrral n es grande (en la práctica, $n \geq 30$ es suficiente), la distribución de los estadísticos D_n, D_n^+ y D_n^- bajo H_0 puede aproximarse según indica la sigiente proposición.

Proposición 1.2 *1. Para $z > 0$*

$$\lim_{n\longrightarrow\infty} P(\sqrt{n}D_n \leq z) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2z^2}.$$

2. Para $z > 0$

$$\lim_{n\longrightarrow\infty} P(\sqrt{n}D_n^+ \leq z) = 1 - e^{-2z^2}.$$

3. Para tamaños muestrales n grandes

$$4n(D_n^+)^2 \approx \chi_2^2.$$

4. Para tamaños muestrales n grandes y $\alpha = 0{,}05$

$$D_{n,\alpha} \approx \frac{1{,}36}{\sqrt{n}},\ D_{n,\alpha}^+ = D_{n,\alpha}^- \approx \frac{1{,}22}{\sqrt{n}}.$$

1.2.3. Bondad de ajuste a un modelo paramétrico.

Se trata de contrastar

$H_0 : F = F_\theta$ para algún $\theta \in \Theta$, frente a $H_1 : F \neq F_\theta$ para ningún $\theta \in \Theta$.

Sea $\hat{\theta}$ el estimador máximo verosímil de θ calculado a partir de la muestra observada. El estadístico del contraste de Kolmogorov-Smirnov queda modificado como sigue:

$$\hat{D}_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_{\hat{\theta}}(x)|.$$

La distribución de este estadístico no coincide con la de D_n. Además esa distribución depende de la familia paramétrica que se especifica en la hipótesis nula. Algunos casos concretos están tabulados (por ejemplo, en el caso de contrastar normalidad este test se conoce como test de Lilliefors).

1.3. Contrastes de localización en una muestra o en dos muestras apareadas

En sta sección nos planteamos contrastar si la mediana de una muestra es un valor dado, y si la diferencia entre los datos de dos muestras tiene mediana igual a 0.

Sea $X_1, \ldots, X_n$ m.a.s. de $X \sim F$. Sea $M = \text{mediana}(F)$, desconocida, y sea M_0 un valor conocido. Se desea contrastar

$H_0 : M = M_0$ frente a $H_1 : M \neq M_0$ (o $H_1 : M > M_0$, o $H_1 : M < M_0$).

En el caso de datos apareados, $(X_1, Y_1), \ldots, (X_n, Y_n)$ es una m.a.s. de (X, Y) y se desea contrastar la hipótesis nula

$$H_0 : M_D = M_0,$$

donde M_D es la mediana de la variable diferencia $D = X - Y$. En este caso el valor M_0 que se contrasta usualmente es $M_0 = 0$.

Ejemplo 1.2

Un grupo de 7 pacientes con temor al vómito siguen un tratamiento que consiste en la exposición repetida a una película de 4 minutos de duración en la que aparecen escenas de gente vomitando. El efecto que se desea obtener con esta terapia es la reducción de la ansiedad causada por el vómito o la sensación de náuseas. Cada paciente pasó un test, antes y después del tratamiento, que evaluaba su sensación de temor (valores altos indican más temor). Los resultados de estos tests están recogidos en la tabla siguiente:

Paciente	Antes (X)	Después (Y)	Diferencia (D)	Signo de D
1	10.60	7.15	3.45	+
2	7.90	9.36	-1.46	-
3	12.40	6.27	6.13	+
4	16.80	7.19	9.61	+
5	13.20	5.45	7.75	+
6	14.70	6.21	8.49	+
7	18.34	8.00	10.34	+

A la vista de los datos, ¿puede afirmarse que el tratamiento tuvo los efectos deseados?
Si el tratamiento no tuviese efectos, se esperaría que las diferencias entre X e Y fuesen positivas o negativas con igual probabilidad ($H_0 : M_D = 0$). Pero vemos que sólo hay 1 diferencia negativa, mientras que 6 son positivas. ¿Es ésta evidencia suficiente para rechazar H_0?
Si H_0 es cierta, ese resultado es tan probable como sacar 6 caras en 7 lanzamientos de una moneda. De hecho la probabilidad de obtener un resultado tan o menos favorable a H_0 es

$$\left(\frac{1}{2}\right)^7 + 7\left(\frac{1}{2}\right)^6 \frac{1}{2} = 0{,}0625,$$

que será el p-valor del contraste de H_0 basado en el número de signos positivos y negativos. Por lo tanto, no podemos rechazar H_0 a nivel $\alpha = 0{,}05$.

.

1.3.1. El test del signo

El ejemplo anterior es una aplicación de un contraste general conocido como el TEST DEL SIGNO, que formalmente es como sigue.

Dada la muestra $X_1, \ldots, X_n$ de $X \sim F$, que se supone ABSOLUTAMENTE CONTINUA y con mediana M, y planteada la hipótesis nula $H_0 : M = M_0$,

se asigna un signo + a cada observación $X_i > M_0$, y un signo - si $X_i < M_0$. Se usa como estadístico del contraste

$$S = \text{número de signos +}.$$

Obsérvese que bajo H_0

$$Y_i = I\{X_i > M_0\} \sim \text{Bernoulli}(p = 0{,}5)$$

y que

$$S = \sum_{i=1}^{n} Y_i \sim B(n, p = 0{,}5),$$

con lo que queda perfectamente determinada la DISTRIBUCIÓN EXACTA del estadístico del contraste para cualquier tamaño muestral. Obsérvese que esta distribución es independiente de la distribución F de los datos: el test del signo es de DISTRIBUCIÓN LIBRE.

Para n grande ($n > 20$ es suficiente) se puede aproximar la distribución exacta de S por la distribución normal de parámetros $\mu = n/2$ y $\sigma^2 = n/4$. Es recomendable usar una corrección por continuidad en esta aproximación asintótica:

$$P(S \leq r) \approx P\left(Z \leq \frac{r - n/2 + 0{,}5}{\sqrt{n/4}}\right),$$

donde $Z \sim N(0, 1)$.

La siguiente tabla resume la forma de llevar a cabo el test del signo:

Hipótesis nula	Hipótesis alternativa	Rechazar H_0 si ...	p-valor
$M = M_0$	$M > M_0$	S_{Obs} grande	$P(S \geq S_{Obs})$
$M = M_0$	$M < M_0$	S_{Obs} pequeño	$P(S \leq S_{Obs})$
$M = M_0$	$M \neq M_0$	S_{Obs} lejos de $n/2$	$2\,\text{mín}\{1/2, P(S \geq S_{Obs}), P(S \leq S_{Obs})\}$

Vale la pena mencionar que el test del signo puede adaptarse trivialmente para contrastar si el cuantil p de la distribución F, al que llamaremos $Q_p(F)$, es igual a un valor dado Q_0 frente a que es ditinto, mayor o menor que Q_0. El caso de la mediana corresponde a $p = 0{,}5$.

Por último señalemos que en el caso (muy improbable, al suponerse F absolutamente continua) de que alguna observación sea igual a M_0, se elimina ésta y se reduce el tamaño muestral n consecuentemente.

1.3.2. Test de Wilcoxon de los rangos signados

El test del signo sólo utiliza la información de si cada dato es mayor o menor que la mediana M_0 propuesta bajo H_0, pero desaprovecha la información relativa a la magnitud de la diferencia entre las observaciones y M_0. El test de Wilcoxon de los rangos signados sí tiene en cuenta esa información. Para poder aplicarlo se requiere una hipótesis adicional: la distribución F de X ha de ser SIMÉTRICA alrededor de su mediana M.

La hipótesis de simetría de X alrededor de su mediana permite reexpresar esta variable como

$$X \sim M + (2Z - 1)A,$$

donde $Z \sim$ Bernoulli$(1/2)$, $A \sim |X - M|$, y Z y A son variables aleatorias independientes. Observar que $(2Z - 1)$ toma los valores 1 y -1 con probabilidades $1/2$.

Dada la muestra $X_1, \ldots, X_n$ de $X \sim F$, que se supone ABSOLUTAMENTE CONTINUA y SIMÉTRICA alrededor de su mediana M, y planteada la hipótesis nula $H_0 : M = M_0$, se descompone la información contenida en cada X_i en dos partes:

- Se asigna un signo `+` a cada observación $X_i > M_0$, y un signo `-` si $X_i < M_0$, como en el test del signo. De forma equivalente se puede definir $Z_i = I\{X_i > M_0\}$.
- Se calcula $A_i = |X_i - M_0|$.

Bajo la hipótesis nula, A_i y Z_i son independientes y, por lo tanto, es como si los signos `+` y `-` se hubiesen asignado aleatoriamente, sin guardar relación con el tamaño de A_i. Por el contrario, si H_0 es falsa (para fijar ideas podemos suponer que $M > M_0$) los signos `+` tenderán a acompañar a valores grandes de A_i y los signos `-` corresponderán a valores pequeños de A_i.

Así, tiene sentido basar el contraste de H_0 en los siguientes estadísticos:

- T^+, definido como la suma de los RANGOS de los A_i a los que corresponden signos `+`.
- T^-, definido como la suma de los RANGOS de los A_i a los que corresponden signos `-`.

En estas definiciones, el RANGO de un valor $A_i = |X_i - M_0|$ es el lugar que ocupa este valor en la lista ordenada de los valores $A_1, \ldots, A_n$. Más formalmente, sea

$$A_{(1)} < A_{(2)} < \cdots < A_{(n)}$$

la muestra $A_1, \ldots, A_n$ ordenada. El rango de A_i es $R(i) = j$ si y sólo si $A_{(j)} = A_i$. Diremos también que $i = R^{-1}(j)$.

Con esta notación,

$$T^+ = \sum_{i=1}^{n} R(i) Z_i = \sum_{j=1}^{n} j Z_{R^{-1}(j)},$$

$$T^- = \sum_{i=1}^{n} R(i)(1 - Z_i) = \sum_{j=1}^{n} j(1 - Z_{R^{-1}(j)}),$$

Observar que la suma de T^+ y T^- es una cantidad fija:

$$T^+ + T^- = \sum_{j=1}^{n} j = \frac{n(n+1)}{2}.$$

Por lo tanto basta usar uno de ellos (T^+ por ejemplo) como estadístico del contraste.

Por otra parte, si $B_1, \ldots, B_n$ es una m.a.s. de una Bernoulli($p = 1/2$), entonces bajo H_0

$$T^+ \sim \sum_{j=1}^{n} j B_j,$$

lo que implica, por un lado, que la distribución de T^+ bajo H_0 no depende de la distribución desconocida F (el contraste basado en T^+ es de DISTRIBUCIÓN LIBRE) y, por el otro, que la distribución exacta de T^+ es perfectamente conocida: es una v.a. discreta que puede tomar los valores naturales del 0 al $n(n+1)/2$, y la probabilidad de que tome uno de esos valores t es

$$P(T^+ = t) = \frac{n(t)}{2^n},$$

donde $n(t)$ es el número de formas de asignar 0 y 1 a las variables $B_1, \ldots, B_n$ de tal forma que $T^+ = T^+(B_1, \ldots, B_n) = t$. El denominador 2^n es el número de asignaciones equiprobables de 0 y 1 a los B_i. Esta distribución está tabulada en muchos libros.

Observar que

$$E(T^+) = \sum_{j=1}^{n} j \frac{1}{2} = \frac{n(n+1)}{4},\ V(T^+) = \sum_{j=1}^{n} j^2 \frac{1}{4} = \frac{n(n+1)(2n+1)}{24}.$$

Para n grande la distribución de T^+ se puede aproximar por una normal con esos valores como esperanza y varianza. Conviene hacer corrección por continuidad.

Si se producen empates en los valores A_i (cosa poco probable, dado que se supone que F es absolutamente continua) se asigna a cada uno de los valores empatados el promedio de los rangos que tendrían si no hubiese habido empates.

La siguiente tabla resume la forma de llevar a cabo el test de los rangos signados de Wilcoxon:

Hipótesis nula	Hipótesis alternativa	Rechazar H_0 si ...	p-valor
$M = M_0$	$M > M_0$	T^+_{Obs} grande	$P(T^+ \geq T^+_{Obs})$
$M = M_0$	$M < M_0$	T^+_{Obs} pequeño	$P(T^+ \leq T^+_{Obs})$
$M = M_0$	$M \neq M_0$	T^+_{Obs} lejos de $n(n+1)/4$	$2\,\text{mín}\{1/2, P(T^+ \geq T^+_{Obs}), P(T^+ \leq T^+_{Obs})\}$

Ejemplo 1.2, página 10. Continuación. En el ejemplo de los pacientes con temor al vómito, contrastaremos

$$H_0 : M_D = 0 \text{ frente a } H_1 : M_D > 0,$$

donde M_D es la mediana de la diferencia de puntuaciones *Antes menos Después* en los tests que miden ansiedad. Hemos visto antes que el test del signo no encontraba evidencia suficiente para rechazar la hipótesis nula en favor de la alternativa. Veamos qué ocurre si aplicamos el test de los rangos signados de Wilcoxon.

La tabla siguiente recoge los cálculos necesarios:

Paciente	Antes (X_i)	Después (Y_i)	$D_i = X_i - Y_i$	Signo de D_i	$A_i = \lvert D_i \rvert$	Rango R_i
1	10.60	7.15	3.45	+	3.45	2
2	7.90	9.36	-1.46	-	1.46	1
3	12.40	6.27	6.13	+	6.13	3
4	16.80	7.19	9.61	+	9.61	6
5	13.20	5.45	7.75	+	7.75	4
6	14.70	6.21	8.49	+	8.49	5
7	18.34	8.00	10.34	+	10.34	7

Así, el estadístico del test de Wilcoxon es

$$T^+_{Obs} = 2 + 3 + 6 + 4 + 5 + 7 = 27,$$

mientras que $T^-_{Obs} = 1$. El p-valor correspondiente a ese valor se determina usando la tabla de la distribución de T^+. Para $n = 7$,

$$P(T^+ \geq 27) = 0{,}016$$

por lo que se rechaza H_0 a nivel $\alpha = 0{,}05$.

Observar que en este caso podemos calcular ese p-valor sin necesidad de recurrir a las tablas:

$$P(T^+ \geq 27) = P(T^+ = 27) + P(T^+ = 28) = \frac{n(27) + n(28)}{2^7} =$$

$$\frac{2}{2^7} = \frac{1}{64} = 0{,}015625 \approx 0{,}016.$$

Hemos usado que $n(27) = n(28) = 1$ puesto que únicamente las configuraciones

$$(- + + + + + + +) \text{ y } (+ + + + + + + +)$$

dan lugar a valores del estadístico T^+ de 27 y 28, respectivamente.

·———————————————————————————————·

1.4. Comparación de dos muestras independientes

En esta sección nos planteamos contrastar si dos variables aleatorias tienen la misma distribución. Sea $X_1, \ldots, X_m$ m.a.s. de $X \sim F_X$ y sea $Y_1, \ldots, Y_n$ m.a.s. de $Y \sim F_Y$, muestras independientes una de otra. Se supone que F_X y F_Y son absolutamente continuas. Se desea contrastar

$$H_0 : F_X(x) = F_Y(x) \text{ para todo } x \in \mathbb{R}$$

frente a

$$H_1 : F_X(x) \neq F_Y(x) \text{ para algún } x \in \mathbb{R},$$

o

$$H_1 : F_X(x) > F_Y(x) \text{ para todo } x \in \mathbb{R},$$

o

$$H_1 : F_X(x) < F_Y(x) \text{ para todo } x \in \mathbb{R}.$$

Veremos en primer lugar un test basado en las funciones de distribución empíricas. Después, se hará la hipótesis adicional de que F_X y F_Y a lo sumo difieren en su mediana y se presentará un contraste de igualdad de medianas.

1.4.1. Test de Kolmogorov-Smirnov para dos muestras

Sean $F_{X,m}$ y $F_{Y,n}$ las funciones de distribución empíricas de ambas muestras. Los estadísticos del contraste de Kolmogorov-Smirnov para dos muestras son éstos:

$$D_{m,n} = \sup_x |F_{X,m}(x) - F_{Y,n}(x)|,$$

$$D^+_{m,n} = \sup_x (F_{X,m}(x) - F_{Y,n}(x)),\ D^-_{m,n} = \sup_x (F_{Y,n}(x) - F_{X,m}(x)).$$

La siguiente tabla resume la forma de llevar a cabo los contrastes de igualdad de distribuciones:

Hipótesis nula	Hipótesis alternativa	Región crítica (nivel α)	p-valor
$F_X(x) = F_Y(x)$	$F_X(x) \neq F_Y(x)$	$D^{Obs}_{m,n} > D_{m,n,\alpha}$	$P(D_{m,n} > D^{Obs}_{m,n})$
$F_X(x) = F_Y(x)$	$F_X(x) > F_Y(x)$	$D^{+,Obs}_{m,n} > D^+_{m,n,\alpha}$	$P(D^+_{m,n} > D^{+,Obs}_{m,n})$
$F_X(x) = F_Y(x)$	$F_X(x) < F_Y(x)$	$D^{-,Obs}_{m,n} > D^-_{m,n,\alpha}$	$P(D^-_{m,n} > D^{-,Obs}_{m,n})$

Para encontrar los valores $D_{m,n,\alpha}$, $D^+_{m,n,\alpha}$ y $D^-_{m,n,\alpha}$ o calcular los p-valores es necesario conocer la distribución de los estadísticos $D_{m,n}$, $D^+_{m,n}$ y $D^-_{m,n}$.

Veamos que bajo H_0, y con $F_X = F_Y = F$, esas distribuciones no dependen de la verdadera distribución F desconocida.

Proposición 1.3 *Supongamos que H_0 es cierta, es decir que $F_X = F_Y = F$, y que F es absolutamente continua. Entonces las distribuciones de $D_{m,n}$, $D^+_{m,n}$ y $D^-_{m,n}$ no dependen de F.*

Demostración: El argumento es análogo al que se siguió en la Proposición 1.1. Razonando como allí, se prueba que

$$F_{X,m}(x) = F^U_m(F(x)),\ F_{Y,n}(x) = F^V_n(F(x)),$$

donde F^U_m es la función de distribución empírica de $U_1, \ldots, U_m$, una m.a.s. de una $U([0,1])$, y F^V_n es la función de distribución empírica de $V_1, \ldots, V_n$, otra m.a.s. de la $U([0,1])$ independiente de la anterior. Así,

$$D_{m,n} = \sup_{x\in\mathbb{R}} |F_{X,m}(x) - F_{Y,n}(x)| = \sup_{x\in\mathbb{R}} |F^U_m(F(x)) - F^V_n(F(x))| =$$

$$\sup_{u\in[0,1]} |F^U_m(u) - F^V_n(u)|,$$

que es el valor del estadístico de Kolmogorov-Smirnov para dos muestras calculado a partir de dos m.a.s. independientes de una $U([0,1])$. Por lo tanto

la distribución de $D_{m,n}$ no depende de F_0. Análogos argumentos pueden hacerse para $D^+_{m,n}$ y $D^-_{m,n}$. □

Las DISTRIBUCIONES EXACTAS de los estadísticos de Kolmogorov-Smirnov para dos muestras se pueden calcular para cualquier tamaño muestral (o pueden aproximarse mediante simulación) y están tabuladas en las referencias indicadas al principio del capítulo.

Obsérvese que la distribución de $D^-_{m,n}$ coincide con la de $D^+_{n,m}$ para cualesquiera tamaños muestrales m y n. También se puede probar que éstas coinciden con las distribuciones de $D^+_{m,n}$ y $D^-_{n,m}$.

En cuanto a la DISTRIBUCIÓN ASINTÓTICA de los estimadores, ésta viene dada en la siguiente proposición.

Proposición 1.4 *1. Para $z > 0$*

$$\lim_{m,n\longrightarrow\infty} P(\sqrt{\frac{mn}{m+n}}D_{m,n} \leq z) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2z^2}.$$

2. Para $z > 0$

$$\lim_{m,n\longrightarrow\infty} P(\sqrt{\frac{mn}{m+n}}D^+_{m,n} \leq z) = 1 - e^{-2z^2}.$$

3. Para tamaños muestrales m, n grandes

$$4\frac{mn}{m+n}(D^+_{m,n})^2 \approx \chi^2_2.$$

4. Para tamaños muestrales n grandes y $\alpha = 0{,}05$

$$D_{m,n,\alpha} \approx 1{,}36\sqrt{\frac{m+n}{mn}},\ D^+_{m,n,\alpha} = D^-_{m,n,\alpha} \approx 1{,}22\sqrt{\frac{m+n}{mn}}.$$

1.4.2. Test de Mann-Whitney-Wilcoxon

Supongamos ahora que las distribuciones de X e Y sólo pueden diferir en su mediana. Es decir,

$$X = M_X + \varepsilon_X,\ Y = M_Y + \varepsilon_Y,$$

donde $\varepsilon_X \sim F_0$, $\varepsilon_Y \sim F_0$ y F_0 es una distribución con mediana 0. En este contexto el contraste de igualdad de distribuciones se reduce a contrastar igualdad de medianas:

$H_0 : M_X = M_Y$ frente a $H_1 : M_X \neq M_Y$ (o $H_1 : M_X > M_Y$, o $H_1 : M_X < M_Y$).

Sean $X_1, \ldots, X_m$ e $Y_1, \ldots, Y_n$ dos m.a.s. independientes de X e Y, respectivamente.

Bajo la hipótesis nula, las $m+n$ observaciones forman una m.a.s. de una única distribución y su etiquetado como una "X" o una "Y" es totalmente aleatorio. Por tanto, si ordenamos los $(m+n)$ datos y les asignamos el rango (posición) correspondiente en la muestra conjunta, la suma T_X de los rangos de las observaciones etiquetadas con "X" (por ejemplo) no será ni muy grande ni muy pequeño si H_0 es cierta, mientras que si realmente $M_X > M_Y$ entonces esta suma T_X tenderá a ser grande, y si $M_X < M_Y$ entonces T_X será en general pequeño.

Ese estadístico T_X es el propuesto por Wilcoxon para contrastar la igualdad de medianas. Más formalmente, T_X es

$$T_X = \sum_{j=1}^{m+n} jI_j,$$

donde

$$I_j = \begin{cases} 1 & \text{si la observación con rango } j \text{ proviene de la muestra de } X, \\ 0 & \text{si la observación con rango } j \text{ proviene de la muestra de } Y. \end{cases}$$

El hecho de basarse en los rangos de los datos hace que su distribución bajo H_0 no dependa de la verdadera distribución, común a todos los datos en ese caso.

La DISTRIBUCIÓN EXACTA puede calcularse para cualesquiera valores m y n y está tabulada en las referencias del capítulo. Se puede probar que

$$E(T_X) = \frac{m(m+n+1)}{2},\ V(T_X) = \frac{mn(m+n+1)}{12}.$$

La DISTRIBUCIÓN ASINTÓTICA de T_X es normal con esos parámetros.

Un estadístico alternativo a T_X fue propuesto por Mann y Whitney. Se trata de

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n} U_{ij},\ \text{donde } U_{ij} = \begin{cases} 1 & \text{si } Y_j < X_i, \\ 0 & \text{si } Y_j > X_i. \end{cases}$$

Se puede probar que

$$U = T_X - m(m+1)/2,$$

por lo que es equivalente basar el contraste en T_X o en U. Por eso el test recibe el nombre de Mann-Whitney-Wilcoxon.

La siguiente tabla resume la forma de llevar a cabo el contraste:

Hipótesis nula	Hipótesis alternativa	Rechazar H_0 si ...	p-valor
$M_X = M_Y$	$M_X > M_Y$	$T_{X,Obs}$ grande	$P(T_X \geq T_{X,Obs})$
$M_X = M_Y$	$M_X < M_Y$	$T_{X,Obs}$ pequeño	$P(T_X \leq T_{X,Obs})$
$M_X = M_Y$	$M_X \neq M_Y$	$T_{X,Obs}$ lejos de $m(m+n+1)/2$	$2\,\text{mín}\{1/2, P(T_X \geq T_{X,Obs}), P(T_X \leq T_{X,Obs})\}$

1.5. Comparación de más de dos muestras

1.5.1. Muestras independientes: Test de Kruskal-Wallis

En esta sección se extiende al caso de más de dos muestras el test de Mann-Whitney-Wilcoxon para comparar medianas en dos muestras independientes.

Sean $X_1, \ldots, X_k$ k variables aleatorias cuyas distribuciones pertenecen a la familia de localización de la distribución F con mediana 0. Es decir,

$$X_j \sim M_j + \varepsilon_j,\ j = 1, \ldots, k,$$

donde $\varepsilon_j \sim F$ y $M_j \in \mathbb{R}$ es la mediana de X_j, para $j = 1, \ldots, k$. Se desea contrastar

$$H_0 : M_1 = \cdots = M_k \text{ frente a } H_1 : \text{No todas las medianas son iguales.}$$

Para ello se observa una m.a.s. de longitud n_j de cada X_j, para $j = 1, \ldots, k$.

Bajo H_0 las $N = n_1 + \cdots + n_k$ observaciones forman una m.a.s. de una única distribución y su pertenencia a una muestra u otra de las k posibles es totalmente aleatoria. Por lo tanto si se asigna a cada observación el rango (posición) que tiene en la muestra de los N datos ordenados de menor a mayor, se tiene que bajo H_0 estos rangos corresponden a cada una de las k muestras de forma aleatoria. Se espera por tanto que los rangos correspondientes a cada muestra estén situados en torno al valor esperado del rango de una observación cualquiera, que es $(N+1)/2$.

Sea R_j la suma de los rangos correspondientes a las n_j observaciones de la muestra j-ésima, y sea $\overline{R}_j = R_j/n_j$ el valor medio de los rangos en esa muestra. Bajo H_0 se espera que $\overline{R}_j$ sea próximo a $(N+1)/2$, para todo j. Por contra, si H_0 es falsa, las muestras correspondientes a poblaciones con mediana M_j pequeñas (respectivamente, grandes) tenderán a concentrar valores bajos (respectivamente, altos) de los rangos. Es decir, si H_0 es falsa $\overline{R}_j$ se situará lejos de $(N+1)/2$ para algunas de las muestras (o quizás para todas ellas).

El estadístico del test de Kruskal-Wallis para contrastar H_0 frente a H_1 es

$$Q = \frac{\sum_{j=1}^{k} n_j \left(\overline{R}_j - \frac{N+1}{2}\right)^2}{\frac{N(N+1)}{12}} = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1).$$

Su DISTRIBUCIÓN EXACTA no depende de la verdadera distribución F que define las v.a. X_j, porque sólo se usa la información de las posiciones que ocupan los datos, y éstas posiciones serían las mismas si transformásemos los datos mediante $F(x)$ para obtener muestras de la $U([0,1])$. El cálculo de esta distribución exacta se lleva a cabo teniedo en cuenta que bajo H_0 la asignación de rangos a los N datos es equivalente a asignarles aleatoriamente una de las $N!$ permutaciones de los números $1, \ldots, N$. La APROXIMACIÓN ASINTÓTICA a esta distribución es ésta:

$$Q \approx \chi^2_{k-1}$$

si $\text{mín}_j\{n_j\}$ es grande.

Obsérvese que el test de Kruskal-Wallis es la versión no paramétrica del contraste de ausencia de efectos en un modelo ANOVA unifactorial.

1.5.2. Muestras relacionadas: Test de Friedman

En esta sección se presenta un test no paramétrico que corresponde al contraste ANOVA de ausencia de efectos de un tratamiento en un diseño por bloques completamente aleatorizado. Es decir, se trata de un diseño con dos factores, uno de los cuales es el factor de interés (el tratamiento, con dos o más niveles) y el otro (el bloque) recoge las diferencias controlables entre los individuos (por ejemplo, diferente instrumental de medida, diferente centro de experimentación, etc.). Nos centraremos en el caso en que haya una única observación por celda (cruce de factores).

Se tienen $N = BT$ observaciones independientes, cada una de ellas con distribución dada por

$$X_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij},\ i = 1, \ldots, B,\ j = 1, \ldots, T,$$

donde las variables ε_{ij} son una m.a.s. de una distribución F absolutamente continua con mediana 0. Se desea contrastar que el tratamiento no tiene efecto:

$$H_0 : \tau_1 = \cdots = \tau_T$$

frente a la alternativa de que no todos los τ_j son iguales.

El test de Friedman utiliza los rangos de las observaciones ordenadas dentro de cada bloque. Bajo H_0 la ordenación de los datos en cada bloque

es una permutación aleatoria de los números $1, \ldots, T$, y por tanto la suma en toda la tabla de los rangos asignados a cada tratamiento tenderá a ser similar para todos los tratamientos. El valor esperado de esta suma es

$$\frac{B\frac{T(T+1)}{2}}{T} = \frac{B(T+1)}{2}.$$

Así, el estadístico del contraste es función de las diferencias entre las sumas de rangos observadas en los tratamientos $R_1, \ldots, R_T$ y $B(T+1)/2$. Concretamente, el estadístico del test de Friedman es

$$S = \frac{12}{BT(T+1)} \sum_{i=1}^{T} \left(R_i - \frac{B(T+1)}{2} \right)^2 = \frac{12}{BT(T+1)} \sum_{i=1}^{T} R_i^2 - 3B(T+1).$$

Su DISTRIBUCIÓN EXACTA bajo H_0 no depende de la distribución F desconocida, ni de los valores μ, β_i o τ_j (iguales estos últimos bajo H_0). El cálculo de la distribución exacta se realiza teniendo en cuenta que cada asignación de rangos es una de las $(T!)^B$ equiprobables. ASINTÓTICAMENTE,

$$S \approx \chi^2_{T-1}$$

si B es grande.

1.6. Medida de la dependencia

PROBLEMA: Sea $(X_1, Y_1), \ldots, (X_n, Y_n)$ una m.a.s. de la v.a. bivariante (X, Y) absolutamente continua, cuya distribución conjunta F_{XY} es desconocida. Se desea cuantificar el *grado de dependencia* de las variables X e Y. También se quiere contrastar

$H_0 : X, Y$ son independientes, frente a $H_1 : X, Y$ no son independientes.

El coeficiente de correlación usual (también denominado *de Pearson*) es la medida de dependencia más usada. En el caso de normalidad conjunta, hay independencia si y sólo si este coeficiente es igual a 0. No obstante, la distribución en el muestreo del coeficiente de correlación de Pearson depende de la verdadera distribución de los datos y es en general desconocida (incluso para el caso de normalidad), lo que fuerza al uso de aproximaciones asintóticas. Por otra parte, el coeficiente de correlación de Pearson no es invariante frente a transformaciones monótonas de los datos.

En esta sección se presentan dos medidas no paramétricas de asociación, cuya distribución bajo la hipótesis de independencia no depende de las distribuciones marginales de X e Y. Además su DISTRIBUCIÓN EXACTA es conocida para todo tamaño muestral n.

1.6.1. Coeficiente τ de Kendall

Sean (X_1, Y_1) y (X_2, Y_2) dos observaciones independientes de (X, Y), v.a. absolutamente continua. Se definen la probabilidad de concordancia como

$$\pi_C = P(X_1 < X_2, Y_1 < Y_2) + P(X_1 > X_2, Y_1 > Y_2) = P((X_1 - X_2)(Y_1 - Y_2) > 0)$$

y la probabilidad de discrepancia como

$$\pi_D = P(X_1 < X_2, Y_1 > Y_2) + P(X_1 > X_2, Y_1 < Y_2) = P((X_1 - X_2)(Y_1 - Y_2) < 0).$$

Por ser (X, Y) absolutamente continua se tiene que

$$\pi_C + \pi_D = 1.$$

En el caso de que X e Y sean independientes se tiene que

$$\pi_C = \pi_D = \frac{1}{2}.$$

Se define el COEFICIENTE τ DE KENDALL (poblacional) como

$$\tau = \pi_C - \pi_D.$$

La letra griega τ se lee *tau*. Este coeficiente tiene las siguientes propiedades:

- $\tau \in [-1, 1]$ y toma los valores 1 o -1 sólo en el caso de relación funcional perfecta y monótona entre X e Y.
- Si X e Y son independientes entonces $\tau = 0$. El recíproco no es cierto en general.
- Si (X, Y) es normal bivariante con coeficiente de correlación de Pearson ρ, entonces
$$\tau = \frac{2}{\pi} \arcsin(\rho).$$
Por lo tanto, bajo normalidad independencia equivale a que τ sea igual a 0.

Se define el COEFICIENTE τ_n DE KENDALL (muestral) como el siguiente estimador insesgado de τ:

$$\tau_n = \frac{1}{\binom{n}{2}} \sum_{i<j} A_{ij},$$

donde $A_{ij} = \text{signo}(X_i - X_j) \times \text{signo}(Y_i - Y_j)$.

Propiedades de τ_n son las siguientes:

1. $\tau_n \in [-1, 1]$ y toma los valores 1 o -1 sólo en el caso de que las dos muestras estén ordenadas de la misma manera.

2. τ_n sólo depende de de los rangos de las observaciones, y no de sus magnitudes.

3. $E(\tau_n) = \tau$.

4. $V(\tau_n) \longrightarrow 0$ cuando $n \longrightarrow \infty$.

5. $\tau_n \longrightarrow \tau$ en probabilidad cuando $n \longrightarrow \infty$.

6. Bajo H_0 (independencia) la DISTRIBUCIÓN EXACTA de τ_n es simétrica y no depende de las distribuciones marginales de X e Y.

7. Bajo H_0 la DISTRIBUCIÓN ASINTÓTICA de τ_n es la siguiente: cuando n tiende a infinito

$$\frac{3\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}\tau_n \longrightarrow N(0,1) \text{ en distribución.}$$

El estadístico τ_n sirve para contrastar independencia:

$$\begin{cases} H_0 : X, Y \text{ son independientes,} \\ H_1 : \tau_{XY} \neq 0 \text{ (o } H_1 : \tau_{XY} < 0, \text{ o } H_1 : \tau_{XY} > 0). \end{cases}$$

La siguiente tabla recoge cómo llevar a cabo el contraste:

Hipótesis nula	Hipótesis alternativa	Rechazar H_0 si ...	p-valor
X e Y indep.	$\tau_{XY} \neq 0$	$\lvert\tau_{n,Obs}\rvert$ grande	$2P(\tau_n \geq \lvert\tau_{n,Obs}\rvert)$
X e Y indep.	$\tau_{XY} > 0$	$\tau_{n,Obs}$ grande	$P(\tau_n \geq \tau_{n,Obs})$
X e Y indep.	$\tau_{XY} < 0$	$\tau_{n,Obs}$ pequeño	$P(\tau_n \leq \tau_{n,Obs})$

1.6.2. Coeficiente de correlación de rangos de Spearman

Sea $(X_1, Y_1), \ldots, (X_n, Y_n)$ una m.a.s. de la v.a. bivariante (X, Y) absolutamente continua. A cada observación X_i le asignamos su rango R_i en la muestra de las X's ordenadas, y a la Y_i le asignamos su rango S_i en la muestra ordenada de las Y's. A partir de ahora trabajaremos con la muestra bivariante de los rangos: $(R_1, S_1), \ldots, (R_n, S_n)$.

El coeficiente de correlación de rangos de Spearman es el coeficiente de correlación usual calculado con las muestras de los rangos (R_i, S_i):

$$R = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{(\sum_{i=1}^{n}(R_i - \bar{R})^2)(\sum_{i=1}^{n}(S_i - \bar{S})^2)}}.$$

Es fácil probar que los valores medios $\bar{R}$ y $\bar{S}$ valen $(n+1)/2$ y que las varianzas del denominador son

$$\sum_{i=1}^{n}(R_i - \bar{R})^2 = \sum_{i=1}^{n}(S_i - \bar{S})^2 = \frac{n(n^2-1)}{12}.$$

Por su parte el numerador se puede escribir como $\sum_{i=1}^{n} R_i S_i - n(n+1)^2/4$. Así, el coeficiente R tiene esta expresión alternativa:

$$\frac{12}{n(n^2-1)}\sum_{i=1}^{n} R_i S_i - 3\frac{n+1}{n-1}.$$

Algunas de las propiedades de R son las siguientes:

1. $R \in [-1, 1]$ y toma los valores 1 o -1 sólo en el caso de que las dos muestras estén ordenadas de la misma manera.

2. R sólo depende de de los rangos de las observaciones, y no de sus magnitudes.

3. Bajo H_0 (independencia) la DISTRIBUCIÓN EXACTA de R es simétrica y no depende de las distribuciones marginales de X e Y.

4. Bajo H_0 se tiene que $E(R) = 0$ y $V(R) = 1/(n-1)$.

5. Bajo H_0 la DISTRIBUCIÓN ASINTÓTICA de R es la siguiente: cuando n tiende a infinito

$$\sqrt{n-1}R \longrightarrow N(0,1) \text{ en distribución.}$$

El estadístico R sirve para contrastar independencia:

$H_0 : X, Y$ son independientes, frente a $H_1 : X, Y$ no son independientes

(o $H_1 : X, Y$ están relacionados positivamente, o $H_1 : X, Y$ están relacionados negativamente.) La siguiente tabla recoge cómo llevar a cabo el contraste:

Hipótesis nula	Hipótesis alternativa	Rechazar H_0 si ...	p-valor
X e Y indep.	X, Y relacionados	$\|R_{Obs}\|$ grande	$2P(R \geq \|R_{Obs}\|)$
X e Y indep.	X, Y relac. posit.	R_{Obs} grande	$P(R \geq R_{Obs})$
X e Y indep.	X, Y relac. negat.	R_{Obs} pequeño	$P(R \leq R_{Obs})$

Señalemos por último que la aproximación de R (estandarizado) a la distribución normal estándar es más lenta que la de τ_n (centrado y estandarizado). Por otro lado, τ_n es un estimador insesgado de la cantidad τ, que tiene una interpretación clara, mientras que no existe ningún parámetro poblacional que sea estimado por el coeficiente de correlación de Spearman R. Estos motivos hacen más atractivo el uso del coeficiente τ_n que el de R.

1.7. Comentarios finales

Empates. Los métodos que hemos visto requieren la hipótesis de continuidad absoluta en la distribución de las variables aleatorias observadas. Por lo tanto, no contemplan la posibilidad de que haya *empates* entre datos, lo cual es relevante especialmente en aquellos que se basan en rangos. En la práctica si el número de empates es pequeño lo que se suele hacer es asignar a los datos empatados el rango promedio que tendrían si no hubiese habido empate (guardando siempre el orden con respecto a las restantes observaciones). De todos modos, existen versiones de los estadísticos que permiten empates entre observaciones y formas de hacer inferencia exacta en estas circunstancias. Ver Gibbons (1993a) y Gibbons (1993b), por ejemplo.

Corrección por continuidad. En las aproximaciones asintóticas de las distribuciones de estadísticos que sólo toman valores naturales es conveniente hacer siempre la corrección por continuidad.

Intervalos de confianza. Sólo hemos visto procedimientos no paramétricos clásicos para contrastar hipótesis. La mayoría de ellos pueden modificarse para dar intervalos de confianza para los parámetros de interés: mediana, diferencia de mediandas o coeficiente τ poblacional. Por ejemplo, un test bilateral para la mediana puede usarse para dar un intervalo de confianza para ésta, definiéndolo como

$$IC_{(1-\alpha)}(M) = \{m \in \mathbb{R} : \text{no se rechaza } H_0 : M = m \text{ a nivel } \alpha\}.$$

Ver Gibbons (1993a) y Gibbons (1993b), por ejemplo.

Comparaciones múltiples de las medianas de más de dos poblaciones. Cuando se rechaza la hipótesis nula de igualdad de medianas en $k \geq 3$ subpoblaciones, siempre es interesante saber qué pares de medianas pueden considerarse iguales y cuáles distintas. Se trata pues de hacer simultáneamente $k(k-1)/2$ contrastes de hipótesis. Estos contrastes individuales se deben hacer a un nivel α^* tal que garantice que la probabilidad de error de Tipo I global (probabilidad de rechazar al menos una hipótesis de igualdad entre dos medianas, cuando la hipótesis nula de igualdad entre todas ellas es cierta) sea igual al nivel α deseado. En Gibbons (1993b) puede verse cómo adaptar los contrastes de Kruskal-Wallis y de Friedman para realizar comparaciones múltiples entre cada par de medianas.

Robustez. Se dice que un procedimento estadístico es robusto frente a la presencia de observaciones atípicas si el resultado obtenido a partir de una muestra no puede ser modificado arbitrariamente mediante la contaminación de la muestra con una proporción pequeña de datos atípicos. Por ejemplo, la media muestral no es robusta, porque alterando un único dato x_i de forma que se le haga tender a infinito, podemos hacer que la media muestral vaya también a infinito. Por contra, la mediana es robusta porque aunque se modificase la mitad menos uno de los datos no conseguiríamos que el valor de la mediana se alejase infinitamente del valor inicial.

Los procedimientos estadísticos basados en el cálculo de momentos (medias, varianza, etc.) suelen presentar problemas de falta de robustez frente a datos atípicos. En cambio, los procedimientos que trabajan con los rangos de las observaciones son robustos. En este sentido los contrastes no paramétricos que hemos estudiado son más robustos que los correspondientes contrastes paramétricos.

Eficiencia relativa. La eficiencia relativa asintótica de un procedimiento de contraste A frente a otro B es el cociente del tamaño muestral que requiere el test B entre el que requiere el test A para obtener ambos contrastes la misma potencia, cuando ambos trabajan al mismo nivel de significación, las hipótesis sobre las distribuciones son las mismas y los tamaños muestrales son grandes. Por ejemplo, la eficiencia relativa asintótica del test del signo frente al test basado en la t de Student es de 0.67 para la distribución normal. Eso significa que el test del signo basado en una muestra de una distribución normal de tamaño 100, por ejemplo, es tan eficiente como el test de la t de Student basado en una de tamaño 67.

La Tabla 1.1 recoge la información sobre eficiencia relativa de los contrastes no paramétricos que hemos estudiado frente a sus competidores paramétricos.

Test no paramétrico	Test paramétrico	ERA bajo normalidad	ERA bajo uniformidad	ERA bajo continuidad y simetría	¿ERA$\geq$ 1?
Signo	t de Student una muestra	0.67		$\geq$0.33	Algunas distribuciones
Rangos signados	t de Student una muestra	0.955	1	$\geq$0.864	Distribuciones con colas más pesadas que la normal
Mann-Whiney-Wilcoxon	t de Student dos muestra	0.955	1	$\geq$0.864	Distribuciones con colas pesadas
Kruskal-Wallis	1-way ANOVA F test	0.955	1	$\geq$0.864	Puede serlo
Friedman	1-way ANOVA F test con T tratamientos y medidas repetidas	$0.955T/(T+1)$	$T/(T+1)$	$\geq 0.864T/(T+1)$	Puede serlo. Por ejemplo, es $3T/(2(T+1))$ para la doble exponencial
τ de Kendall	Coef. Corr. Pearson	0.912			Puede serlo
Coef. Corr. Spearman	Coef. Corr. Pearson	0.912			Puede serlo

Cuadro 1.1: Eficiencia Relativa Asintótica (ERA) de los contrastes no paramétricos frente a los correspondientes tests paramétricos.

Capítulo 2

Introducción a los métodos de estimación no paramétrica de curvas

REFERENCIAS: Capítulo 1 de Simonoff (1996), Algunos ejemplos de los capítulos 1, 2 y 3 de Fan y Gijbels (1996). Otras referencias: Bowman y Azzalini (1997). Silverman (1986), Wand y Jones (1995), Wasserman (2006)

2.1. Introducción

LOS MÉTODOS DE ESTIMACIÓN NO PARAMÉTRICA DE CURVAS, también conocidos como MÉTODOS DE SUAVIZADO, son una colección de técnicas estadísticas que tienen por objetivo estimar una curva relacionada con la distribución de probabilidad de los datos observados, haciendo las menos hipótesis posibles.

Las funciones que habitualmente se estiman son:

- la *función de densidad*, sus derivadas o su integral (*función de distribución*):

$$X \sim f(x),\ F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du,\ f(x) = F'(x).$$

- la *función de regresión* o sus derivadas:

$$(X, Y) \sim F(x, y),\ m(x) = E(Y|X = x).$$

- la *función de riesgo*, sus derivadas o su integral (*función de riesgo acumulada*):

$$X \sim f(x),\ \lambda(x) = \frac{f(x)}{1 - F(x)},\ \Lambda(x) = \int_{-\infty}^{x} \lambda(u)du,\ \lambda(x) = \Lambda'(x).$$

- la *curva principal*, que es una versión no lineal de la primera componente principal.

Ejemplo 2.1
Estimación de la densidad.

En el Capítulo 1 de Simonoff (1996) se presenta el siguiente ejemplo. Se trata de datos sobre el interés que pagan 69 entidades financieras en uno de sus productos llamado Certificados de Depósito. El conjunto de datos se conoce como CD rate data. El fichero `cdrate.dat` (ver la pàgina web que acompaña al libro Simonoff 1996) contiene esos datos junto con una variable binaria que indica si las entidades son bancos (0) o cajas de ahorros (1).

Una primera representación de los datos es mediante un diagrama de tallo y hojas (Stem-and-Leaf Plot):

```
The decimal point is 1 digit(s) to the left of the |

74 | 167
76 | 15
78 | 2200
80 | 0000000000556157
82 | 0550003334556
84 | 00000005999000000000001257
86 | 550158
```

Este gráfico permite visualizar la distribución de los datos (es como un histograma girado) sin perder información del valor numérico de los datos.

Una mejor representación gráfica la obtenemos mediante un histograma de los datos, tal como se recoge en el primer panel de la Figura 2.1. El histograma es un estimador no paramétrico de la función de densidad. Muestra qué zonas de la recta acumulan más probabilidad y cuáles menos. En este ejemplo se aprecia bimodalidad en los datos. El histograma tiene un inconvenienete fundamental: es una función poco suave (de hecho es discontinua en los bordes de cada una de las cajas) y es constante a trozos. Estas características no son las que acostumbran a tener las funciones de densidad. Otra

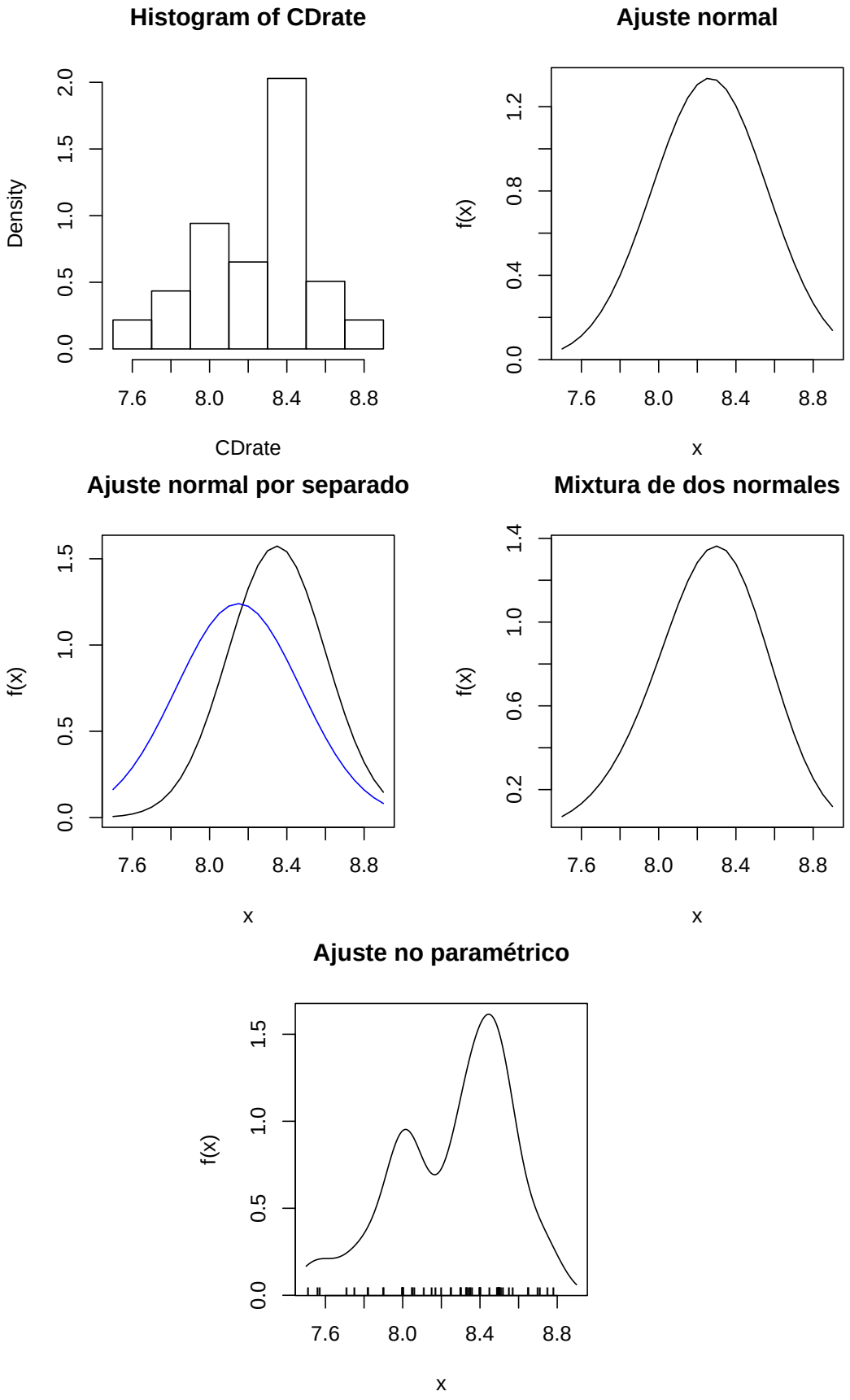


Figura 2.1: Diferentes estimaciones paramétricas y no paramétricas de la densidad de la variable CDrate.

forma de tener un estimador de la densidad de la variable `CDrate` es hacer una hipótesis paramétrica. Por ejemplo podemos suponer que esta variable es normal y estimar sus parámetros mediante la media y la desviación típica muestrales (ver segundo panel de la Figura 2.1). Esta alternativa también tiene serios inconvenientes: el modelo paramétrico no se ajusta bien a los datos porque es excesivamente rígido (por ejemplo, el modelo normal impone unimodalidad y simetría, lo que no es acorde con el histograma de los datos).

Dado que el histograma sugiere bimodalidad, podemos pensar que ésta se debe a que los datos provienen de mezclar dos poblaciones, bancos y cajas, con distribuciones quizás diferentes. Ello nos hace pensar que un posible modelo es la mixtura de dos normales. La segunda fila de gráficos de la Figura 2.1 muestra los resultados de este ajuste: a la izquierda se muestran las dos densidades normales ajustadas en cada subpoblación y a la derecha la mixtura de ambas. Se ha corregido la falta de asimetría del ajuste con una única normal, pero sigue habiendo unimodalidad.

Un estimador no paramétrico de la densidad alternativo al histograma es el estimador núcleo (ver Capítulo 3). Este estimador aplicado a los datos de `CDrate` da como resultado la densidad representada en el último gráfico de la Figura 2.1. Este estimador es suave y respeta la bimodalidad y la asimetría de los datos.

Ejemplo 2.2

Regresión con respuesta continua.

Consideremos el siguiente ejemplo, en el que se analiza la relación entre dos variables del conjunto de datos referido a la vivienda en 506 barrios de Boston en 1978 (Boston Housing Data; ver por ejemplo
`http://lib.stat.cmu.edu/datasets/boston_corrected.txt`, o
`http://www.ailab.si/orange/doc/datasets/housing.htm`).
Concretamente, se busca expresar la variable `room` (número medio de habitaciones por vivienda) como función de la variable `lstat` (porcentaje de población con estatus social en la categoría inferior). Para ello podemos utilizar un modelo de regresión lineal que ajuste la variable `room` como función de la variable `lstat`. El resultado se muestra en el panel superior izquierdo de la Figura 2.2. Se observa que el patrón lineal impuesto por el modelo paramétrico elegido es muy rígido para adaptarse a la relación existente entre las variables. Esto se debe a que la relación entre las variables no es lineal: la variable room desciende bruscamente cuando la variable lstat pasa del 0 %

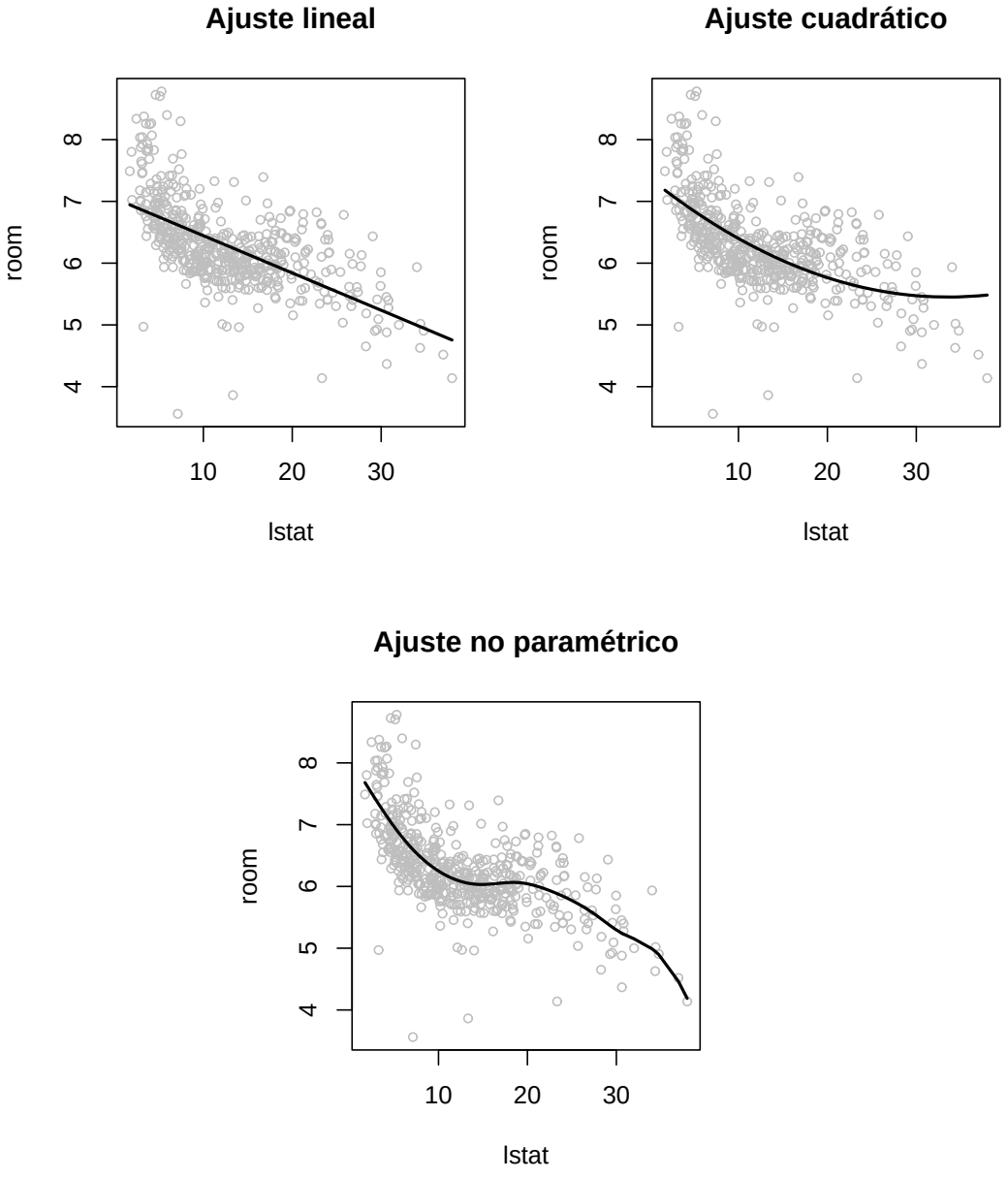


Figura 2.2: Ajustes paramétricos y no paramétrico de la variable `room` como función de la variable `lstat`.

al 10 %, pero después se mantiene prácticamente constante. Una posible solución sería introducir un término cuadrático (el cuadrado de `lstat`) en el modelo de regresión para reflejar la ausencia de linealidad. El segundo panel de la Figura 2.2 muestra el resultado. Pero, aún así, el nuevo modelo de regresión paramétrico no consigue adaptarse totalmente a los datos observados.

Por último, realizamos el ajuste no paramétrico de la variable `room` como función de la variable lstat (en el Capítulo 4 se explicará con detalle cómo se obtiene este ajuste). El panel inferior de la Figura 2.2 muestra los resultados de este ajuste no paramétrico. La relación entre las variables es distinta según si el porcentaje de población de extracción social más baja (`lstat`) es inferior al 10 %, está entre el 10 % y el 20 %, o supera ese valor. En el tramo intermedio, el número medio de habitaciones por vivienda (`room`) se mantiene constante, mientras que en los otros tramos decrece al crecer `lstat`. Además, la disminución es más acusada en el primer tramo.

Este ejemplo muestra que la modelización no paramétrica es mucho más flexible que la paramétrica y permite resaltar los patrones de dependencia presentes en los datos.

.———————————————————————————————————.

Ejemplo 2.3

Regresión con respuesta binaria.

En el siguiente ejemplo, extraído de Fan y Gijbels (1996), se relaciona la probabilidad de supervivencia tras sufrir quemaduras de tercer grado con la superficie de piel afectada. Son datos referidos a 435 adultos (entre 18 y 85 años de edad) que fueron tratados por quemaduras de tercer grado en el Centro de Quemados del Hospital General de la University of Southern California. Para cada paciente, se tienen dos variables: `lgae`, el logaritmo de 1 más el área de las quemaduras de tercer grado, y `superv`, una variable 0-1 que indica si el paciente sobrevive o no. Observar que se trata de un problema de regresión binaria, en el que la esperanza condicionada de `superv` dado un nivel de `lgae` es la probabilidad condicionada de sobrevivir condicionada a ese valor de `lgae`. En la Figura 2.3 se muestran los datos y las probabilidades de supervivencia ajustadas con un modelo logístico y con un método no paramétrico (la función menos suave).

.———————————————————————————————————.

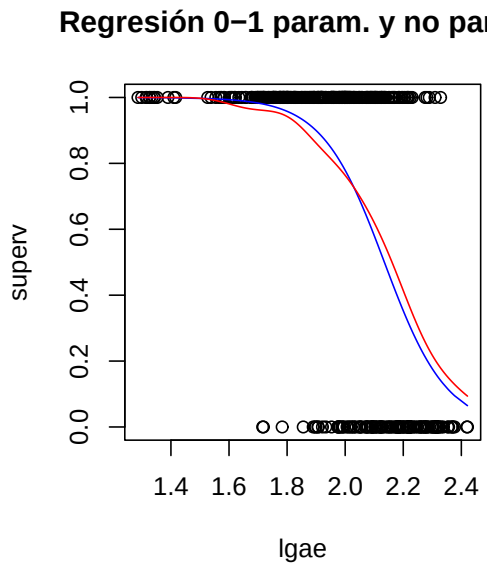


Figura 2.3: Ajustes paramétrico y no paramétrico de la probabilidad de supervivencia como función de la variable `lgae`.

Ejemplo 2.4
Curvas principales.
Dado un conjunto de datos con una configuración no elíptica, se busca la curva que mejor ajusta los datos. Las curvas principales fueron definidas por Hastie and Stuetzle (1989) como *curvas parametrizadas suaves que atraviesan una nube de puntos multidimensional por su parte central*. Son generalizaciones no lineales de la primera componente principal. La Figura 2.4 muestra un conjunto de datos de dimensión 3 que ha sido generado añadiendo ruido a posiciones aleatorias sobre una espiral unidimensional. También se muestra la curva principal estimada a partir de los datos.

2.2. Usos de los métodos de suavizado.

Análisis exploratorio de datos. Permiten obtener gráficos de la *función de densidad*, la *función de regresión*, la *función de riesgo* o sus derivadas (entre otras). En análisis multivariante permiten tener versiones

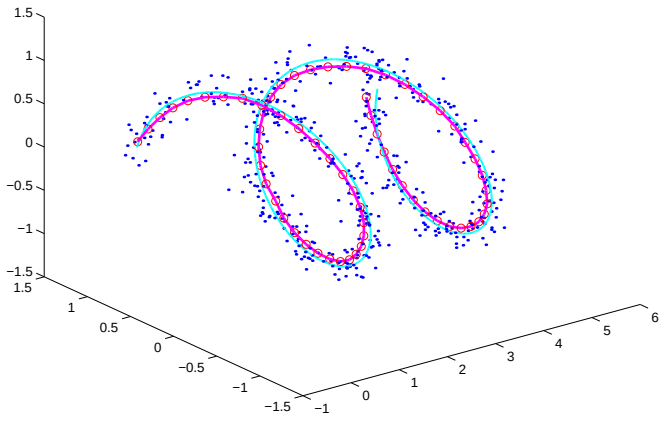


Figura 2.4: Curva principal ajustada a un conjunto de datos artificiales.

no lineales y no paramétricas de las componentes principales (*curvas principales*).

Construcción de modelos. A partir de una descripción fiel de los datos pueden proponerse modelos que se ajusten bien a los datos. Por ejemplo, si la densidad estimada resulta bimodal, podemos proponer una mixtura de dos subpoblaciones como modelo para nuestros datos.

Bondad de ajuste de un modelo paramétrico. Sea $X \sim f$. Se trata de contrastar

$$H_0 : f \in \mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}, \text{ frente a } H_1 : f \notin \mathcal{F}_\Theta.$$

Un estadístico útil para este contraste será

$$T = d(f_{\hat{\theta}}, \hat{f}),$$

donde $\hat{\theta}$ es un estimador de θ (y por lo tanto $f_{\hat{\theta}}$ es un estimador paramétrico de f), $\hat{f}$ es un estimador no paramétrico de f y $d(\cdot, \cdot)$ es una distancia entre funciones de densidad.

Estimación paramétrica. Si suponemos que $X \sim f_{\theta_0}$, con $\theta_0 \in \Theta$, un estimador de θ puede venir dado como

$$\hat{\theta} = \arg \min_{\theta \in \Theta} T = d(f_\theta, \hat{f}).$$

Este tipo de estimadores se llaman de MÍNIMA DISTANCIA.

Definir nuevos métodos estadísticos. En muchos casos es posible modificar una metodología paramétrica estándar sustituyendo $f_{\hat{\theta}}$ por $\hat{f}$. Como ejemplo, veamos el caso del análisis discriminante.

El planteamiento general es que una población se encuentra subdivida en dos subpoblaciones, C_1 y C_2, y que una variable aleatoria X (de dimensión $k \geq 1$) observable en los individuas de esta población tiene distribución distinta en C_1 y C_2:

$$X|C_1 \sim f_1,\ X|C_2 \sim f_2.$$

Se observan los valores de X en n_1 individuos de C_1 y en n_2 de C_2. En base a esa información hay que definir una REGLA DISCRIMINANTE que permita asignar a C_1 o a C_2 un nuevo individuo del que sólo se sabe que presenta un valor x de la variable X.

La Regla Disciminante Lineal de Fisher se basa en el supuesto de normalidad con igual varianza en C_1 y C_2:

$$X|C_1 \sim N(\mu_1, \sigma^2); X|C_2 \sim N(\mu_2, \sigma^2).$$

La forma de proceder es la siguiente: se estiman μ_1 y μ_2 a partir de las muestras de cada subpoblación. Se estima σ^2 conjuntamente a partir de los $(n_1 + n_2)$ datos. La regla discriminante es ésta:

Clasificar el individuo con valor x de X en C_1 si y sólo si

$$f_{(\hat{\mu}_1, \hat{\sigma}^2)}(x) \geq f_{(\hat{\mu}_2, \hat{\sigma}^2)}(x).$$

Se puede probar que esta regla equivale a calcular una función lineal de x y clasificar la nueva observación en C_1 si y sólo si esa función lineal es positiva.

Esta regla se puede modificar para obtener una REGLA DISCRIMINANTE NO PARAMÉTRICA:

Clasificar el individuo con valor x de X en C_1 si y sólo si

$$\hat{f}_1(x) \geq \hat{f}_2(x),$$

donde $\hat{f}_1(x)$ y $\hat{f}_2(x)$ son estimadores no paramétricos de las densidades f_1 y f_2, respectivamente, calculados a partir de las dos muestras observadas.

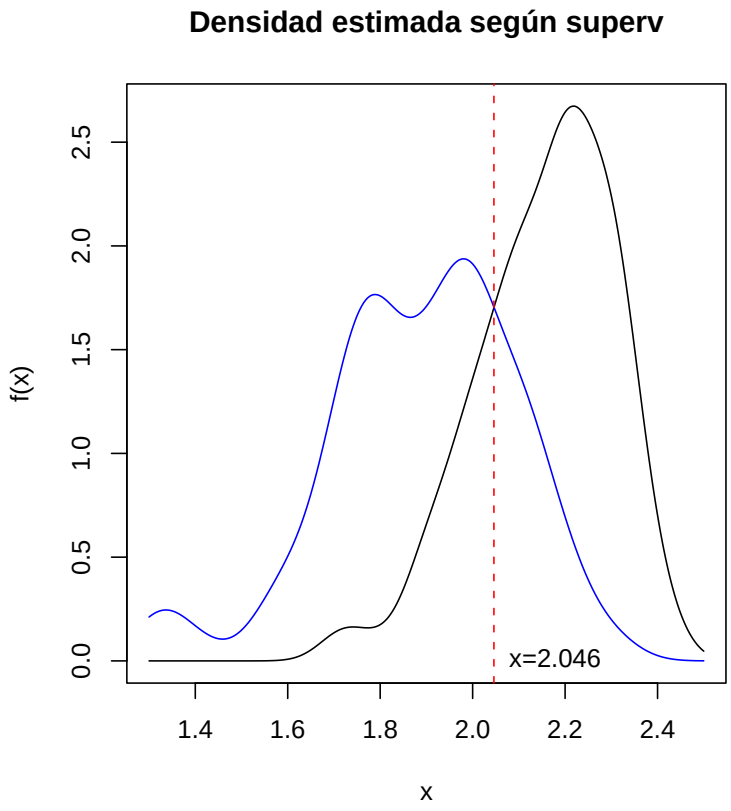


Figura 2.5: Estimación de la densidad de `lgae` en las dos subpoblaciones definidas por `superv`. La correspondiente a los no supervivientes es la densidad que tiene la moda más a la derecha.

Ejemplo 2.5

Consideremos de nuevo el ejemplo en que se relaciona la probabilidad de supervivencia tras sufrir quemaduras de tercer grado con la superficie de piel afectada. Se aplica ahí la regla discriminante no paramétrica basada en la estimación de las densidades en cada subpoblación. El resultado (ver Figura 2.5) es clasificar como potenciales supervivientes a los enfermos con valor de la variable `lgae` menor o igual que 2.046.

Capítulo 3

Estimación no paramétrica de la densidad

REFERENCIAS: Silverman (1986), Scott (1992), Wand y Jones (1995), Simonoff (1996), Fan y Gijbels (1996), Bowman y Azzalini (1997), Wasserman (2006).

3.1. La estimación de la densidad

Sean $x_1, \ldots, x_n$ observaciones independientes de una variable aleatoria X que tiene función de densidad f. Sea $x \in \mathbb{R}$. Se quiere estimar el valor de la función de densidad f en x: $f(x)$. Recordemos algunas propiedades conocidas de la función de densidad:

- Una función de densidad es cualquier función que verifica

$$f(x) \geq 0 \text{ para todo } x \in \mathbb{R}, \quad \int_{-\infty}^{\infty} f(x)dx = 1.$$

- f es función de densidad de X si y sólo si para todo a, b reales con $a \leq b$ se tiene que

$$P(a < X \leq b) = \int_a^b f(x)dx.$$

- Si dx es una longitud pequeña,

$$f(x) \approx \frac{P(X \in [x, x + dx])}{dx}.$$

- Sea $F(x)$ la función de distribución de X. Entonces,

$$F(x) = \int_{-\infty}^{x} f(u)du, \;\; f(x) = \frac{d}{dx}F(x) = F'(x).$$

Una forma de estimar $f(x)$ es hacer supuestos paramétricos sobre la distribución de X:

$$f \in \{f(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^k\}.$$

Por ejemplo se podría suponer que $X \sim N(\mu, \sigma^2)$, y así $k = 2$, $\theta = (\mu, \sigma^2)$. Bajo ese supuesto paramétrico, se usa la muestra observada para estimar el parámetro θ mediante $\hat{\theta}$ (por ejemplo por máxima verosimilitud) y se toma como estimador de $f(x)$ el valor

$$\hat{f}_\theta(x) = f(x;\hat{\theta}).$$

Este procedimiento, que se conoce como *estimación paramétrica de la densidad*, es muy dependiente del modelo elegido. No tiene flexibilidad para detectar desviaciones de esa hipótesis.

Aquí vamos a abordar la *estimación no paramétrica de la densidad*.

3.2. El histograma y el polígono de frecuencias

El primer estimador no paramétrico de la densidad y quizás aún el más utilizado es el HISTOGRAMA. Se construye de la siguiente forma.

Se eligen *marcas* $b_0 < b_1 < \ldots < b_m$ en $\mathbb{R}$ con

$$b_0 < \min_{i=1\ldots n} x_i, \;\; \max_{i=1\ldots n} x_i \leq b_m$$

y se definen los intervalos $B_j = (b_{j-1}, b_j]$, $j = 1, \ldots, m$. Sea n_j el número de observaciones que caen en B_j, y f_j la frecuencia relativa de este intervalo (la proporción de observaciones que caen en B_j):

$$n_j = \#\{x_i : x_i \in B_j\} = \sum_{i=1}^{n} I_{B_j}(x_i), \;\; f_j = \frac{n_j}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{B_j}(x_i).$$

Sobre cada intervalo B_j se dibuja un rectángulo que tiene B_j por base y cuya altura a_j es tal que el área es igual a f_j:

$$a_j(b_j - b_{j-1}) = f_j = \frac{1}{n}\sum_{i=1}^{n} I_{B_j}(x_i) \Longrightarrow a_j = \frac{f_j}{b_j - b_{j-1}}.$$

Sea x el punto donde se quiere estimar la densidad $f(x)$. Si x no está dentro de ningún intervalo B_j el estimador histograma de $f(x)$ es 0. Si $x \in B_j$, el estimador histograma de $f(x)$ es la altura a_j del histograma en B_j:

$$\hat{f}_H(x) = \frac{f_j}{b_j - b_{j-1}} \text{ si } x \in B_j.$$

Observar que la función $\hat{f}_H(x)$ cumple las propiedades de una función de densidad: es no negativa e integra 1.

Usualmente se toman todos los intervalos de la misma anchura: $b_j - b_{j-1} = b$, $j = 1, \ldots, m$. Así $\hat{f}_H(x) = f_j/b$ si $x \in B_j$, lo que también podemos escribir como

$$\hat{f}_H(x) = \sum_{j=1}^{m} \frac{f_j}{b} I_{B_j}(x) = \sum_{j=1}^{m} \frac{n_j}{nb} I_{B_j}(x) = \sum_{j=1}^{m} f_j \frac{1}{b} I_{B_j}(x).$$

Observar que esta última expresión corresponde a la mixtura de m densidades, cada una de ellas uniforme en B_j, con pesos iguales a las frecuencias relativas de cada intervalo B_j.

3.2.1. Motivación del histograma como estimador de la densidad

Recordemos que

$$f(x) = \frac{d}{dx} F(x) = \lim_{h \longrightarrow 0} \frac{F(x+h) - F(x)}{h} = \lim_{u+v \longrightarrow 0} \frac{F(x+u) - F(x-v)}{u+v},\ u \geq 0, v \geq 0.$$

Si dividimos $\mathbb{R}$ en intervalos de amplitud b, con b pequeño, y llamamos a los extremos de los intervalos b_j, $j \in \mathrm{Z}$, un punto $x \in \mathbb{R}$ pertenecerá a uno de esos intervalos: $x \in (b_j, b_{j+1}]$. Sean

$$u = b_{j+1} - x,\ v = x - b_j \Longrightarrow x + u = b_{j+1}, x - v = b_j, u + v = b_{j+1} - b_j = b.$$

Así, si b es pequeño,

$$f(x) \approx \frac{F(x+u) - F(x-v)}{u+v} = \frac{F(b_{j+1}) - F(b_j)}{b}.$$

Si se estima la función de distribución F mediante la función de distribución empírica $\hat{F}$, se tiene que

$$\hat{F}(b_j) = \frac{\#\{x_i \leq b_j\}}{n}.$$

Si se sustituye en la expresión anterior de $f(x)$, obtendremos el siguiente estimador:

$$\begin{aligned}\hat{f}(x) &= \frac{\hat{F}(b_{j+1}) - \hat{F}(b_j)}{b} \\ &= \frac{\#\{x_i \leq b_{j+1}\} - \#\{x_i \leq b_j\}}{nb} \\ &= \frac{\#\{b_j < x_i \leq b_{j+1}\}}{nb} \\ &= \frac{n_j}{nb} = \frac{f_j}{b} = \hat{f}_H(x).\end{aligned}$$

Es decir, llegamos a la expresión del histograma que ya conocíamos.

3.2.2. Características del histograma

1. El histograma es muy simple, tanto de cálculo como de interpretación.
2. Su aspecto depende mucho de la anchura de los intervalos: b.

Ejemplo 3.1

Consideremos el conjunto de datos referido a la vivienda en 506 barrios de Boston (BOSTON HOUSING DATA), que ya fue tratado en el ejemplo 2.2 del Capítulo 2. En la Figura 3.1 se muestran tres histogramas de la variable `LSTAT` (porcentaje de población con estatus social en la categoría inferior). Se han usado anchuras de intervalos b distintas, y el aspecto que presentan los histogramas es bastante distinto. Por ejemplo, el segundo de ellos muestra multimodalidad, mientras que esto pasa desapercibido en el primero.

.

3. El aspecto del histograma depende del ANCLA del histograma, que es el punto desde donde arranca el primer intervalo.

Ejemplo 3.2

La Figura 3.2 muestra la importancia del ancla del histograma. Se ha usado el conjunto de datos relativos a tipos de interés en Certificados de Depósito (ver Ejemplo 2.1, Capítulo 2). La variable representada es `CDrate`.

.

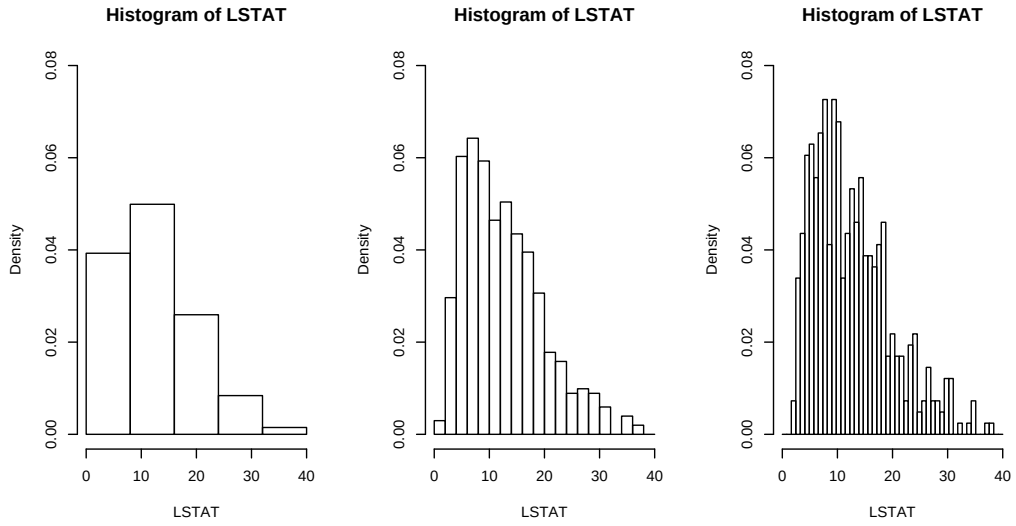


Figura 3.1: Histogramas de la variable `LSTAT` con distintos valores de la anchura de intervalo b.

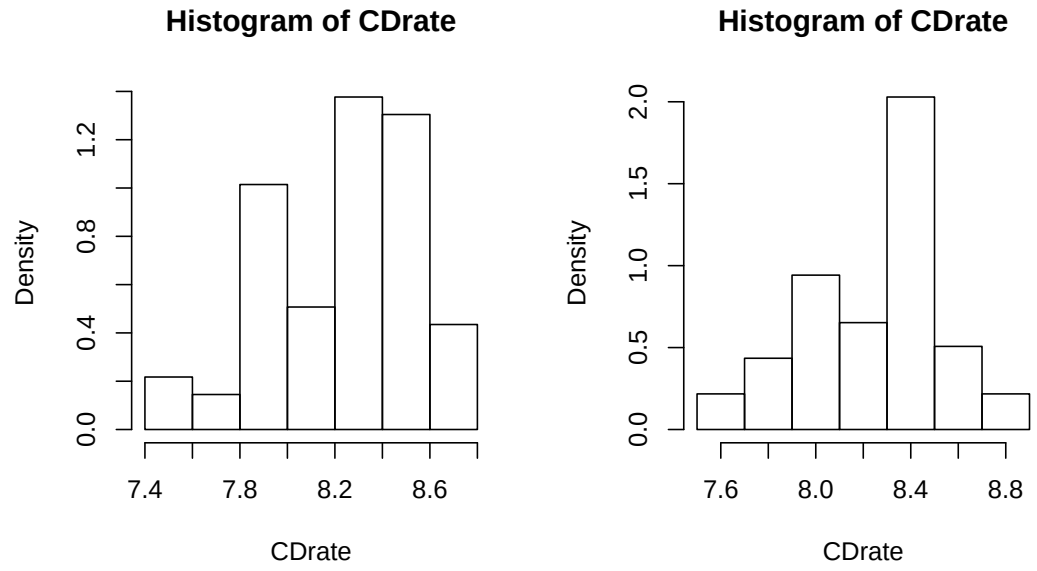


Figura 3.2: Ilustración de la importancia del *ancla* del histograma. La variable representada es `CDrate`.

4. El histograma no es un estimador *suave* de la función de densidad: es discontinuo y constante a intervalos.

5. La anchura de las cajas b tiene una influencia importantísima en el comportamiento del histograma como estimador de la función de densidad (además de la ya mencionada influencia en su aspecto), que puede resumirse como sigue:

   - Si b es pequeño, el histograma tiene poco sesgo y mucha varianza.
   - Si b es grande, el histograma tiene mucho sesgo y poca varianza.

Ejemplo 3.3

La Figura 3.3 muestra los histogramas estimados para muestras de tamaño 100 simuladas a partir de una mixtura de normales con densidad

$$f(x) = \frac{3}{4} f_N(x; \mu = 0, \sigma = 1) + \frac{1}{4} f_N(x; \mu = 3/2, \sigma = 1/3),$$

donde $f_N(x; \mu, \sigma)$ es la función de densidad de una $N(\mu, \sigma^2)$. Esta función aparece como ejemplo en el segundo capítulo de Wand y Jones (1995). En el panel de la izquierda se representan, para 30 muestras, sus histogramas construidos con anchura b de intervalos igual a 2 (b *grande*), mientras que en el de la derecha se usa $b = 0{,}5$ (b pequeño) para otras 30 muestras. Se observa que a la izquierda el sesgo es grande y la varianza pequeña, mientras que lo contrario ocurre a la derecha.

.

3.2.3. Propiedades locales del estimador histograma

En esta sección nos ocuparemos de las propiedades asintóticas del histograma evaluado en un punto x fijo, $\hat{f}_H(x)$, como estimador del valor desconocido $f(x)$. Utilizaremos algunas de las herramientas que se incluyen en el Apéndice (página 187).

Como criterio para evaluar localmente un estimador de $f(x)$ usaremos el Error Cuadrático Medio:

$$\mathrm{MSE}(\hat{f}_H(x)) = E[(\hat{f}_H(x) - f(x))^2] = (\mathrm{Sesgo}(\hat{f}_H(x)))^2 + V(\hat{f}_H(x)).$$

Teorema 3.1 *Supongamos que la anchura $b = b_n$ de los intervalos del histograma decrece hacia 0 cuando n tiende a infinito ($b = o(1)$). Si f tiene segunda derivada continua y acotada, entonces*

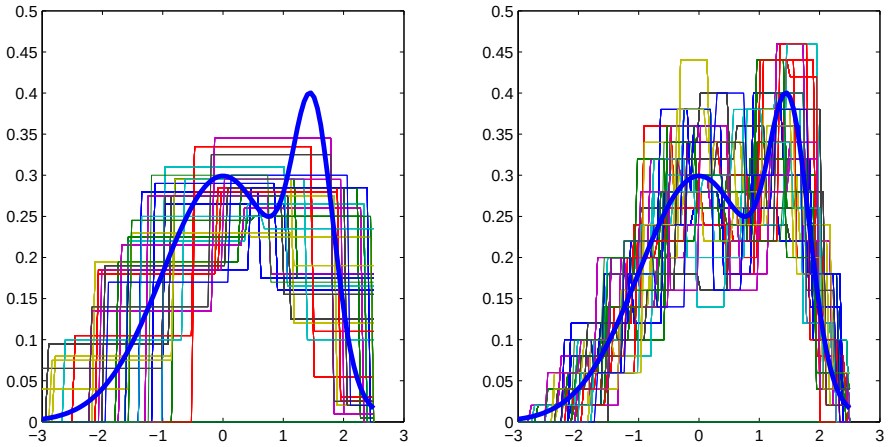


Figura 3.3: Influencia de b en el sesgo y varianza del histograma como estimador de la función de densidad.

1. *si* $x \in (b_j, b_j + b]$,

$$E(\hat{f}_H(x)) = f(x) + \frac{1}{2}f'(x)(b - 2(x - b_j)) + O(b^2) = f(x) + O(b),$$

2.

$$V(\hat{f}_H(x)) = \frac{f(x)}{nb} + O\left(\frac{1}{n}\right),$$

3. *si* $x \in (b_j, b_j + b]$,

$$MSE(\hat{f}_H(x)) = \frac{f(x)}{nb} + \frac{1}{4}(f'(x))^2(b - 2(x - b_j))^2 + O\left(\frac{1}{n}\right) + O(b^3),$$

4. *Si* $b = b_n \longrightarrow 0$ *y* $nb_n \longrightarrow \infty$ *cuando* $n \longrightarrow \infty$ *(es decir,* b_n *tiende a 0 pero no demasiado deprisa) entonces*

$$\hat{f}_H(x) \longrightarrow f(x) \text{ en probabilidad cuando } n \longrightarrow \infty.$$

Demostración: Sea $X_1, \ldots, X_n$ m.a.s. de $X \sim f$ la muestra a partir de la cual construimos el histograma. Hemos visto que

$$\hat{f}_H(x) = \frac{n_j}{nb} \text{ si } x \in B_j = (b_j, b_j + b],$$

donde n_j es el número de observaciones $X_1, \ldots, X_n$ que caen en $(b_j, b_j + b]$. Por lo tanto,

$$n_j \sim B(n, p_j = F(b_j + b) - F(b_j)) \Rightarrow E(n_j) = np_j,\ V(n_j) = np_j(1 - p_j).$$

Así

$$E(\hat{f}_H(x)) = \frac{p_j}{b},\ V(\hat{f}_H(x)) = \frac{p_j(1-p_j)}{nb^2} = \frac{p_j}{nb^2} - \frac{p_j^2}{nb^2}.$$

Por el Teorema de Taylor,

$$F(b_{j+1}) = F(x) + f(x)(b_{j+1} - x) + \frac{1}{2}f'(x)(b_{j+1} - x)^2 + O(b^3),$$

$$F(b_j) = F(x) + f(x)(b_j - x) + \frac{1}{2}f'(x)(b_j - x)^2 + O(b^3).$$

Si restamos esas dos expresiones obtenemos que

$$p_j = F(b_{j+1}) - F(b_j) = f(x)b + \frac{1}{2}f'(x)\left((b + (b_j - x))^2 - (b_j - x)^2\right) + O(b^3)$$

$$= f(x)b + \frac{1}{2}f'(x)(b^2 - 2b(x - b_j)) + O(b^3).$$

Así,

$$E(\hat{f}_H(x)) = f(x) + \frac{1}{2}f'(x)(b - 2(x - b_j)) + O(b^2).$$

Estudiemos ahora la varianza de $\hat{f}_H(x)$. Observar que

$$\frac{p_j}{nb^2} = \frac{f(x)}{nb} + \frac{1}{2}f'(x)\frac{1}{n} - \frac{f'(x)(x - b_j)}{nb} + O\left(\frac{b^2}{n}\right).$$

Como $b = o(1)$ cuando n tiende a infinito $O(b^2/n) = o(1/n)$. Teniendo en cuenta además que $(x - b_j) = O(b)$, se sigue que

$$\frac{p_j}{nb^2} = \frac{f(x)}{nb} + O\left(\frac{1}{n}\right).$$

Por otro lado,

$$\frac{p_j^2}{nb^2} = nb^2\left(\frac{p_j}{nb^2}\right)^2 = nb^2 O\left(\frac{1}{n^2b^2}\right) = O\left(\frac{1}{n}\right).$$

Así se tiene que

$$V(\hat{f}_H(x)) = \frac{f(x)}{nb} + O\left(\frac{1}{n}\right).$$

El resto del enunciado se sigue por argumentos estándares. □

El Teorema anterior muestra que la convergencia del estimador histograma es más rápida en los puntos centrales $c_j = (b_j + b_{j-1})/2$ de los intervalos B_j que en los restantes puntos: el término principal de la expresión asintótica del sesgo

$$\text{Sesgo}(\hat{f}_H(x)) = \frac{1}{2}f'(x)(b - 2(x - b_j)) + O(b^2) = f'(x)(b/2 - (x - b_j)) + O(b^2),$$

se anula en c_j. Así que $\text{Sesgo}(\hat{f}_H(c_j)) = O(b^2)$, mientras que en general ese sesgo es $O(b)$, y

$$\text{MSE}(\hat{f}_H(c_j)) = \frac{f(c_j)}{nb} + O\left(\frac{1}{n}\right) + O(b^4),$$

cuando en general

$$\text{MSE}(\hat{f}_H(x)) = \frac{f(x)}{nb} + O\left(\frac{1}{n}\right) + O(b^2).$$

Ese buen comportamiento del histograma en los puntos centrales de las cajas motivará más adelante las definiciones del POLÍGONO DE FRECUENCIAS (Sección 3.2.6) y del ESTIMADOR NÚCLEO DE LA DENSIDAD (Sección 3.3).

3.2.4. Propiedades globales del estimador histograma

Ahora nos ocuparemos del comportamiento de la *función histograma* $\hat{f}_H$ como estimador de la *función de densidad* f. Una forma de medir la distancia entre estimador y función estimada es integrar el error cuadrático sobre todo el soporte de f, que supondremos que es un intervalo $I \subset \mathbb{R}$ acotado. Se tiene así lo que se conoce como ERROR CUADRÁTICO INTEGRADO (*Integrated Square Error*, en inglés):

$$\text{ISE}(\hat{f}_H) = \int_I (\hat{f}_H(x) - f(x))^2 dx,$$

que es una variable aleatoria porque depende de la muestra $X_1, \ldots, X_n$ de X observada. Su valor esperado (respecto a las muestras $X_1, \ldots, X_n$) es el ERROR CUADRÁTICO INTEGRADO MEDIO (*Mean Integrated Square Error*, en inglés):

$$\text{MISE}(\hat{f}_H) = E[\text{ISE}(\hat{f}_H)] = E\left[\int_I (\hat{f}_H(x) - f(x))^2 dx\right].$$

Observar que

$$\begin{aligned}
\mathrm{MISE}(\hat{f}_H) &= \int \left(\int_I (\hat{f}_H(x) - f(x))^2 dx \right) dF_{X_1,\ldots,X_n} \\
&\quad \text{(el Teorema de Fubini permite cambiar el orden de integración)} \\
&= \int_I \left(\int (\hat{f}_H(x) - f(x))^2 dF_{X_1,\ldots,X_n} \right) dx \\
&= \int_I \mathrm{MSE}(\hat{f}_H(x)) dx = \mathrm{IMSE}(\hat{f}_H).
\end{aligned}$$

Por lo tanto el MISE, que es un promedio del error global, es igual al IMSE (Error Cuadrático Medio Integrado, *Integrated Mean Square Error*, en inglés), que es una medida del error puntual acumulado.

Teorema 3.2 *Para el estimador histograma*

$$MISE(\hat{f}_H) = \frac{1}{nb} + \frac{b^2}{12} R(f') + O\left(\frac{1}{n}\right) + o(b^2),$$

donde para una función g definida en $I \subseteq \mathbb{R}$, $R(\phi) = \int_I \phi(x)^2 dx$.

Demostración: Tenemos en cuenta que $\mathrm{MISE}(\hat{f}_H) = \mathrm{IMSE}(\hat{f}_H)$, y para cada $x \in I$ llamamos $j(x)$ al entero j tal que $x \in B_j$. Así,

$$\begin{aligned}
\mathrm{IMSE}(\hat{f}_H) &= \int_I \left(\frac{f(x)}{nb} + \frac{1}{4}(f'(x))^2(b - 2(x - b_{j(x)}))^2 + O\left(\frac{1}{n}\right) + O(b^3) \right) dx \\
&= \frac{1}{nb} + \sum_{j=1}^{m} \int_{b_j}^{b_{j+1}} (f'(x))^2 \left(\frac{b}{2} - (x - b_j) \right)^2 dx + O\left(\frac{1}{n}\right) + O(b^3) \\
&\quad \text{(por el Teorema del Valor Medio Integral Generalizado,} \\
&\quad \text{y haciendo } u = x - b_j) \\
&= \frac{1}{nb} + \sum_{j=1}^{m} (f'(\psi_j))^2 \int_0^b \left(\frac{b}{2} - u \right)^2 du + O\left(\frac{1}{n}\right) + O(b^3) \\
&\quad \left\{ \int_0^b \left(\frac{b}{2} - u \right)^2 du = \left(-\frac{1}{3} \left(\frac{b}{2} - u \right)^3 \right) \Big|_0^b = \frac{1}{3}\frac{b^3}{8} + \frac{1}{3}\frac{b^3}{8} = \frac{b^3}{12} \right\} \\
&= \frac{1}{nb} + \frac{b^2}{12} \sum_{j=1}^{m} (f'(\psi_j))^2 b + O\left(\frac{1}{n}\right) + O(b^3) \\
&\quad \{\text{la suma de Riemann } \sum_{j=1}^{m} (f'(\psi_j))^2 b = \int_I (f'(x))^2 dx + o(1), \\
&\quad \text{donde } o(1) \longrightarrow 0 \text{ si } b \longrightarrow 0\} \\
&= \frac{1}{nb} + \frac{b^2}{12} \int_I (f'(x))^2 dx + O\left(\frac{1}{n}\right) + o(b^2).
\end{aligned}$$

□

3.2.5. Elección del parámetro de suavizado b

A los términos principales de la expresión asintótica del MISE se les llama AMISE (*Asymptotic Mean Integrated Square Error*, en inglés). En el caso del histograma es

$$\mathrm{AMISE}(\hat{f}_H) = \frac{1}{nb} + \frac{b^2}{12} R(f').$$

El primer sumando $(1/nb)$ proviene de la integral sobre I de la varianza del histograma, mientras que el segundo se debe a la integral del cuadrado del sesgo. Observar el comportamiento de ambos términos en función del ancho b de las cajas del histograma:

- El término de la varianza $(1/nb)$ es decreciente en b.

- El término del sesgo (proporcional a b^2) crece con b.

Por lo tanto la correcta elección del PARÁMETRO DE SUAVIZADO b nos permite buscar un compromiso entre sesgo y varianza del estimador de f. De hecho, podemos buscar el valor de b que minimiza AMISE$(\hat{f}_H)$, al que como función de b podemos denotar por $g(b)$:

$$g(b) = \frac{1}{nb} + \frac{b^2}{12} R(f').$$

Derivamos g respecto a b,

$$g'(b) = -\frac{1}{nb^2} + \frac{2b}{12} R(f')$$

e igualamos a 0,

$$g'(b_0) = 0 \Longleftrightarrow b_0^3 = \frac{6}{nR(f')} \Longleftrightarrow b_0 = \left(\frac{6}{R(f')}\right)^{1/3} n^{-1/3}.$$

Para ese valor b_0 óptimo el AMISE toma el valor

$$\text{AMISE}_0 = n^{-2/3} \left(\frac{6}{R(f')}\right)^{1/3} + n^{-2/3} R(f')^{1/3} \frac{6^{1/3}}{12} = \left(\frac{9}{16} R(f')\right)^{1/3} n^{-2/3}.$$

El problema con la fórmula del parámetro de suavizado óptimo,

$$b_0 = \left(\frac{6}{R(f')}\right)^{1/3} n^{-1/3},$$

es que $R(f') = \int_I f'(x)^2 dx$ es desconocido porque depende de la densidad desconocida f que pretendemos estimar.

La forma más sencilla de superar ese problema es calcular el valor que tendría $R(f')$ si f perteneciese a un modelo paramétrico. Concretamente, si f fuese la densidad de una $N(\mu, \sigma^2)$ entonces

$$R(f') = \frac{1}{4\sqrt{\pi}\sigma^3}.$$

Tomando éste como verdadero valor de $R(f')$, el valor de b_0 es

$$b_0^* = (24\sqrt{\pi})^{1/3} \sigma n^{-1/3} = 3{,}491 \sigma n^{-1/3}.$$

Esta forma de elegir b_0 se conoce como REGLA DE REFERENCIA A LA NORMAL (*normal reference rule*, en inglés).

El valor de σ se estima a partir de los datos observados mediante

$$\hat{\sigma} = \text{mín}\{S, \text{IQR}/1{,}35\},$$

donde S^2 es la varianza muestral de los datos, y IQR es su rango intercuartílico (recordar que el rango intercuartílico de una $N(\mu, \sigma^2)$ es 1,35). Se toma el mínimo de esos dos estimadores naturales de σ para que el estimador final sea más robusto frente a la presencia de datos atípicos.

3.2.6. El polígono de frecuencias

Una de las malas propiedades del histograma es que el estimador de la función f a que da lugar es un función discontinua. El POLÍGONO DE FRECUENCIAS corrige ese defecto. Se define como el interpolador lineal de los valores del histograma en los puntos centrales de cada intervalo,

$$(c_j, \hat{f}_H(c_j)), j = 0, \ldots, m+1,$$

donde $c_0 = b_0 - (b/2)$, $c_{m+1} = b_m + (b/2)$, $\hat{f}_H(c_0) = \hat{f}_H(c_{m+1}) = 0$.

Ejemplo 3.4

Consideremos de nuevo el conjunto de datos referido a la vivienda en 506 barrios de Boston. La Figura 3.4 muestra el polígono de frecuencias de la variable `LSTAT` (porcentaje de población con estatus social en la categoría inferior) construido a partir del histograma que también se muestra en trazo discontinuo.

.

Las siguientes son expresiones alternativas del polígono de frecuencias. Si $x \in [c_j, c_{j+1}]$,

$$\begin{aligned}
\hat{f}_{PF}(x) &= \frac{f_j}{b} + (x - c_j)(f_{j+1} - f_j)\frac{1}{b^2} \\
&= \frac{1}{b}\left(f_j \frac{c_{j+1} - x}{b} + f_{j+1}\frac{x - c_j}{b}\right) \\
&= \frac{1}{b^2}\left(f_j c_{j+1} - f_{j+1}c_j + (f_{j+1} - f_j)x\right).
\end{aligned}$$

El polígono de frecuencias presenta ventajas respecto al histograma, la más clara de las cuales es que proporciona un estimador continuo de la función de densidad. Su comportamiento asintótico también es mejor que el del

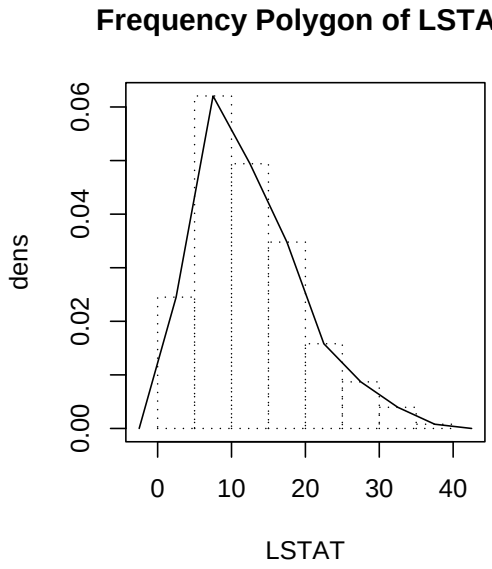


Figura 3.4: Polígono de frecuencias de la variable `LSTAT`.

histograma, como se verá en la sección siguiente. No obstante, la falta de suavidad del polígono de frecuencias (no es derivable en los puntos c_j) hace recomendable buscar otros estimadores de la densidad (ver Sección 3.3).

3.2.7. Comportamiento asintótico del polígono de frecuencias

En la Sección 3.2.3 se ha visto que la convergencia del estimador histograma es más rápida en los puntos centrales $c_j = (b_j + b_{j-1})/2$ de los intervalos B_j que en los restantes puntos. Para construir el polígono de frecuencias sólo se evalúa el histograma en esos puntos centrales c_j, en los que el sesgo puntual del histograma es $O(b^2)$, en vez del orden $O(b)$ general. Se puede probar que esa mejora se mantiene al hacer la interpolación lineal y que, por tanto, el polígono de frecuencias converge más rápidamente que el histograma al verdadero valor de la densidad en todos los puntos. Supongamos que la densidad f tiene como soporte al intervalo acotado $I \subset \mathbb{R}$, que f es derivable 3 veces y que $R(f)$, $R(f')$, $R(f'')$ y $R(f''')$ son finitas (recordar que $R(\phi) = \int_I \phi(x)^2 dx$). La siguientes son las propiedades asintóticas más relevantes del polígono de frecuencias.

Error cuadrático medio integrado:

$$\mathrm{MISE}(\hat{f}_{PF}) = \frac{2}{3nb} + \frac{49b^4 R(f'')}{2880} + O\left(\frac{1}{n}\right) + O(b^6).$$

Parámetro de suavizado óptimo: El ancho de los intervalos b que minimiza el AMISE es

$$b_0 = 2\left(\frac{15}{49R(f'')}\right)^{1/5} n^{-1/5}.$$

AMISE para b_0 : Para ese valor b_0 óptimo el AMISE del polígono de frecuencias toma el valor

$$\mathrm{AMISE}_0 = \frac{5}{12}\left(\frac{49}{15}R(f'')\right)^{1/5} n^{-4/5}.$$

Observar que $\mathrm{AMISE}_0 = O(n^{-4/5})$, lo que es una mejora respecto al AMISE del histograma, que es $O(n^{-2/3})$. El AMISE que tendríamos con un estimador paramétrico (suponiendo que el modelo fuese correcto) sería $O(n^{-1})$. Con estimadores no paramétricos nunca puede alcanzarse esa velocidad de convergencia.

Regla de referencia a la normal: Si f fuese la densidad de una $N(\mu, \sigma^2)$ el ancho óptimo de los intervalos para el polígono de frecuencias sería

$$b_0^* = 2{,}15\sigma n^{-1/5}.$$

3.3. Estimador núcleo de la densidad

El estimador no paramétrico de la función de densidad más usado, exceptuando el histograma, es el ESTIMADOR NÚCLEO. Este estimador introduce dos mejoras destacadas respecto al estimador histograma:

Localización. En la Sección 3.2.3 se ha visto que el histograma es mejor estimador en el centro de cada intervalo que en otros puntos. Hagamos pues que x, el punto donde queremos estimar la densidad, sea el centro de uno de los intervalos del histograma: $B_x = [x - b/2, x + b/2] = [x - h, x + h]$. (Pasar de intervalos semi-abiertos a intervalos cerrados no tiene implicaciones ni teóricas ni prácticas). Así, el estimador de $f(x)$ será

$$\hat{f}_U(x) = \frac{1}{nb}\sum_{i=1}^{n} I_{[x-b/2,x+b/2]}(x_i) = \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{2} I_{[-1,1]}\left(\frac{x - x_i}{h}\right).$$

Cuando se desea estimar la densidad en otro punto x', se sitúa el intervalo del histograma alrededor de x' y se aplica la fórmula anterior. Cuando x recorre $\mathbb{R}$, la función $\hat{f}_U(x)$ así construida constituye un estimador de f. La Figura 3.5 muestra esta estimación de la densidad en el caso de una mixtura de normales, $f(x) = \frac{3}{4}f_N(x; \mu = 0, \sigma = 1) + \frac{1}{4}f_N(x; \mu = 3/2, \sigma = 1/3)$, a partir de una muestra simulada de tamaño 100. Se ha usado $h = 0{,}15$.

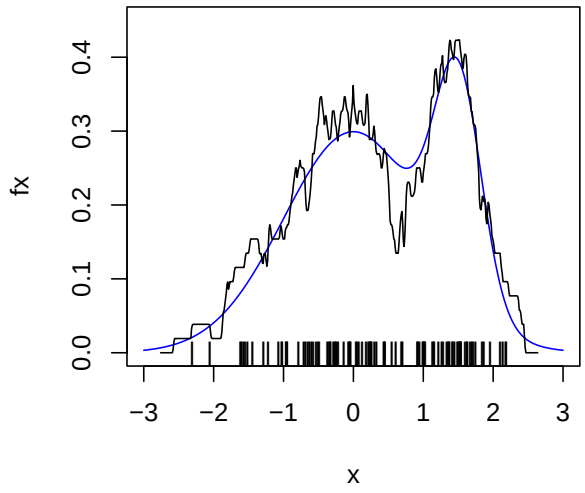


Figura 3.5: Estimación de la densidad mediante un histograma móvil ($h = 0{,}15$) o, dicho de otro modo, usando un estimador núcleo con kernel uniforme.

Suavidad. La función $\hat{f}_U(x)$ anterior no es suave (es discontinua y constante a trozos). La razón es que en su expresión aparece la función de densidad de la v.a. $U([-1, 1])$,

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

que es discontinua y constante a trozos. Si se sustituye esa densidad por otra $K(u)$ más suave (por ejemplo, derivable unas cuantas veces) se obtiene un estimador de la densidad que hereda esas propiedades de suavidad. El estimador resultante

$$\hat{f}_K(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}K\left(\frac{x - x_i}{h}\right) \tag{3.1}$$

se denomina ESTIMADOR NÚCLEO o estimador KERNEL.

La función K se llama FUNCIÓN NÚCLEO (o KERNEL) y, en general, es una función de densidad continua, unimodal y simétrica alrededor del 0. El parámetro h se conoce como PARÁMETRO DE SUAVIZADO.

Otra forma de interpretar el estimador núcleo es observar que es la densidad de la CONVOLUCIÓN de la distribución empírica y la distribución con densidad $K_h(e) = K(e/h)/h$. En efecto, consideremos una v.a. X_K que se construye de la siguiente forma:

1. Generar un ruido ϵ de una v.a. con densidad $K(e/h)/h$.
2. Elegir al azar con equiprobabilidad uno de los n puntos observados $x_1, \ldots, x_n$. Sea x_E el valor elegido.
3. Hacer $X_K = x_E + \epsilon$

Entonces la v.a. X_K tiene función de densidad igual a $\hat{f}_K(x)$. Este estimador distribuye el peso $1/n$ de cada dato observado en un entorno suyo de forma continua, tal como se ilustra en la Figura 3.6, donde hay cinco observaciones, marcadas en la parte inferior.

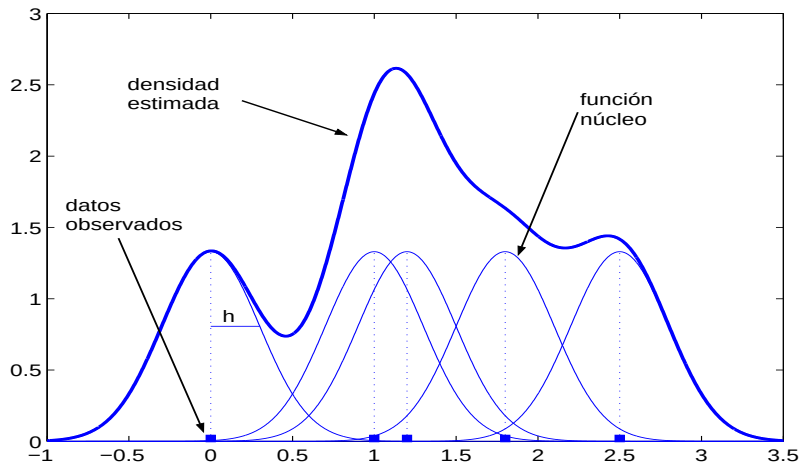


Figura 3.6: Estimación de la función de densidad a partir de cinco observaciones mediante un núcleo gaussiano.

Observar que el estimador núcleo puede expresarse como

$$\hat{f}_K(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - x_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n}K_h(x - x_i).$$

Es decir, $\hat{f}_K$ es la mixtura de n densidades (con pesos $1/n$) con la misma forma que el núcleo K, reescaladas según el parámetro h, y centradas cada una en observación x_i, como se ve en la Figura 3.6.

De todo lo anterior se deduce que el estimador núcleo es una función de densidad (siempre que lo sea K).

El parámetro de suavizado h (o *ventana* o *bandwidth*) controla la concentración del peso $1/n$ alrededor de cada x_i: si h es pequeño únicamente las observaciones x_i más cercanas a x serán relevantes en la estimación de $f(x)$, mientras que valores grandes de h permiten que observaciones más alejadas de x también intervengan en la estimación $\hat{f}(x)$. La Figura 3.7 ilustra el efecto del parámetro h en la apariencia del estimador núcleo.

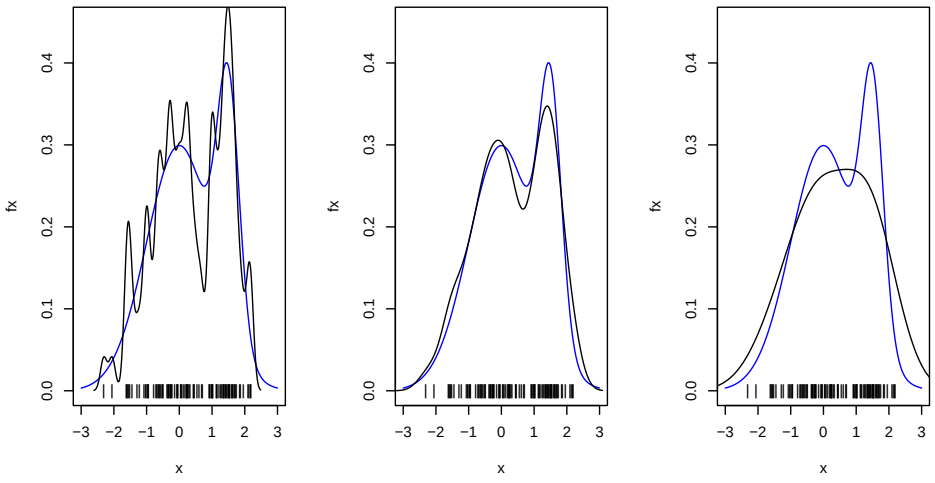


Figura 3.7: Efecto del parámetro h en la apariencia del estimador núcleo en la estimación de una mixtura de dos normales. Los valores de h son 0.1, 0.3 y 0.7, de derecha a izquierda.

La estimación final se ve notablemente afectada por cambios en la elección del parámetro de suavizado, por lo que esta tarea resulta crucial en la estimación no paramétrica de la densidad (en la sección 3.4 se tratará en detalle este problema). Valores grandes de h hacen que los estimadores de la densidad sean muy estables de muestra a muestra (**poca varianza**) pero las estimaciones presentan **gran sesgo**. Por el contrario, si h es pequeño el estimador varía mucho en muestras diferentes (**mucha varianza**), pero en promedio estima bien la densidad desconocida (**poco sesgo**).

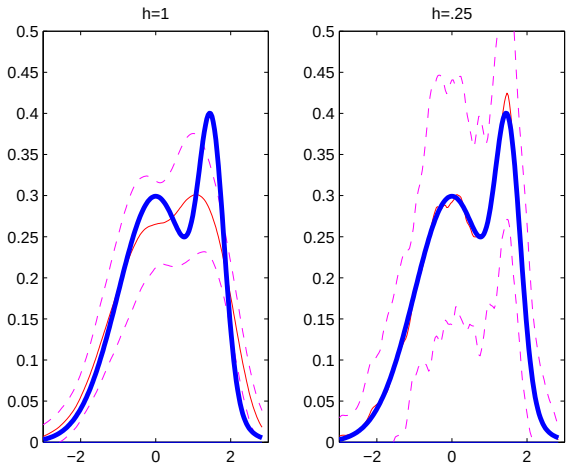


Figura 3.8: Influencia de h en el sesgo y varianza del estimador núcleo de la densidad.

Ejemplo 3.5

La Figura 3.8 muestra un gráfico de variabilidad de 30 estimadores núcleo (el gráfico muestra el promedio de los estimadores y las bandas puntuales situadas a ± 2 desviaciones típicas) construidos a partir de muestras de tamaño 100 simuladas a partir de una mixtura de normales con densidad

$$f(x) = \frac{3}{4} f_N(x; \mu = 0, \sigma = 1) + \frac{1}{4} f_N(x; \mu = 3/2, \sigma = 1/3),$$

donde $f_N(x; \mu, \sigma)$ es la función de densidad de una $N(\mu, \sigma^2)$. En el panel de la izquierda se ha utilizado $h = 1$ (h *grande*), mientras que en el de la derecha se usa $h = 0{,}25$ (h pequeño). Se observa que a la derecha el sesgo es grande y la varianza pequeña, mientras que lo contrario ocurre a la izquierda. Se ha usado un núcleo *biweight* (ver Cuadro 3.1).

Hay una serie de propiedades que hacen que una función K que cumpla algunas de ellas sea una función núcleo (o *kernel*) satisfactoria para ser utilizada en la definición (3.1) del estimador núcleo de la densidad.

1. K **es simétrica alrededor de 0.**
   Es una propiedad deseable, pero no imprescindible. Implica que el peso $1/n$ de cada dato observado se reparte de forma simétrica alrededor de la observación.

2. K **es unimodal** (con moda en 0, si K es además simétrica).
   Es una propiedad deseable, pero no imprescindible. Implica que el peso $1/n$ de cada dato observado se reparte de forma que queda más peso en las zonas más cercanas a la observación.

3. K **es una función de densidad:** $K(u) \geq 0$ para todo $u \in \mathbb{R}$ y $\int_{\mathbb{R}} K(u)du = 1$.
   Esta propiedad garantiza que el estimador núcleo definido en (3.1) es una función de densidad. No es una propiedad necesaria para que el estimador núcleo tenga buenas propiedades asintóticas.

4. K **es positiva:** $K(u) \geq 0$ para todo $u \in \mathbb{R}$.
   No es una propiedad necesaria para que el estimador núcleo tenga buenas propiedades asintóticas.

5. K **integra 1:** $\int_{\mathbb{R}} K(u)du = 1$.
   Ésta es una propiedad necesaria para que el sesgo asintótico del estimador sea nulo.

6. K **tiene momento de orden 1 nulo:** $\int_{\mathbb{R}} uK(u)du = 0$.
   Se cumple si K es simétrica (y tiene esperanza). Si K no tiene esperanza 0 entonces el sesgo del estimador decrece más lentamente hacia 0.

7. K **tiene momento de orden 2 finito:** $\int_{\mathbb{R}} u^2K(u)du = \sigma_K^2 < \infty$.
   Que la varianza de K sea finita es necesario para que el estimador tenga sesgo asintótico acotado. Por otra parte, se pueden construir núcleos no positivos con momento de orden 2 nulo que permiten reducir el sesgo asintótico (son los llamados *núcleos de alto orden*).

8. K **es una función suave** (tiene r derivadas continuas).
   El estimador núcleo hereda las propiedades de suavidad del núcleo K a partir del que se define. Por tanto, es necesario utilizar núcleos suaves para obtener estimadores suaves.

9. K **tiene soporte compacto.**
   Esta propiedad es deseable desde el punto de vista computacional. Si $K(u)$ se anula fuera del intervalo $[-c, c]$, entonces para evaluar $\hat{f}_K$ en un punto x sólo hay que utilizar los puntos x_i situados en $[x - ch, x + ch]$.

3.3.1. Comportamiento asintótico del estimador núcleo de la densidad

Comencemos por recordar la definición de CONVOLUCIÓN DE DOS DENSIDADES.

Definición 3.1 *Sean $X \sim f$, $Y \sim g$ dos variables aleatorias independientes. La convolución $f * g$ de las densidades f y g es la función de densidad de $X + Y$ y vale*

$$(f * g)(x) = \int_{\mathbb{R}} f(x - y)g(y)dy$$

Si $X \sim f$ y $\varepsilon \sim K_h$ son variables aleatorias independientes, con $V(X)$ mucho mayor que $V(\varepsilon)$, la convolución $f * K_h$ de las densidades f y K_h es la función de densidad de $X + \varepsilon$. La densidad $f * K_h$ es un suavizado de la densidad f (un *difuminado* de f) en el que se suavizan picos y valles.

Ejemplo 3.6

Consideremos f la densidad de la mixtura de 4 normales con medias situadas en -3, -1, 1 y 3, y desviación típica 0.5 común a todas ellas. Sea K_h la densidad de una normal centrada en 0 y con desviación típica 0.5. Es fácil comprobar que la convolución $f * K_h$ corresponde a la mixtura de 4 normales con medias situadas en -3, -1, 1 y 3, y desviación típica común igual a $1/\sqrt{2}$. Por lo tanto los marcados picos y los valles de f quedan atenuados en la densidad $f * K_h$. Véase la figura 3.9.

.

Consideremos ahora el problema de estimación no paramétrica de la densidad: $x_1, \ldots x_n$ son n observaciones independientes de la v.a. X que tiene función de densidad desconocida $f(x)$. Sea

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

el estimador núcleo de $f(x)$.

Teorema 3.3 (Comportamiento local del estimador núcleo) *Se suponen las siguientes hipótesis de regularidad:*

1. *$f(x)$ es función con 3 derivadas continuas de x.*
2. *K es simétrica, $\int_{\mathbb{R}} K(u)du = 1$, $\int_{\mathbb{R}} uK(u)du = 0$ y $\int_{\mathbb{R}} u^2K(u)du < \infty$.*

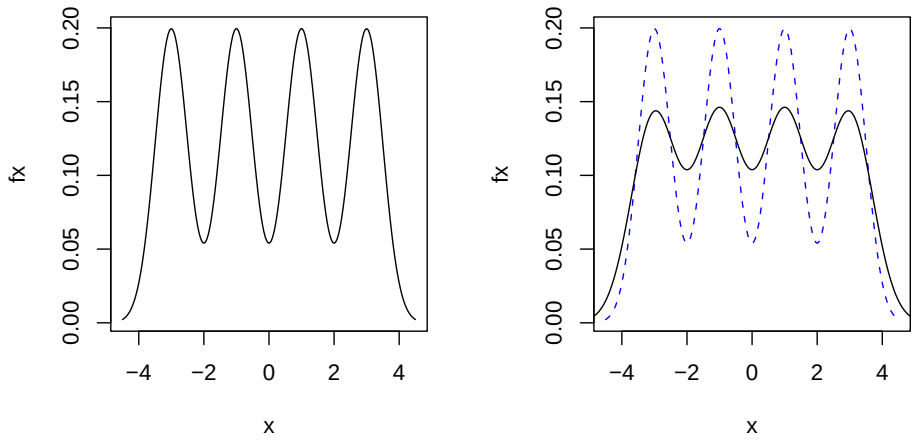


Figura 3.9: Convolución. El gráfico de la derecha muestra la convolución de la densidad de la izquierda con la densidad de un ruido normal con desviación típica 0.5.

3. $(x-h, x+h)$ *está contenido en el soporte de* $f(x)$.

4. $h \longrightarrow 0$ *y* $nh \longrightarrow \infty$ *cuando* $n \longrightarrow \infty$.

El sesgo y la varianza asintóticos de $\hat{f}(x)$ *son de la siguiente forma:*

$$Sesgo(\hat{f}(x)) = E(\hat{f}(x)) - f(x) = (f * K_h)(x) - f(x) = \frac{f''(x)\sigma_K^2 h^2}{2} + O(h^3).$$

$$V(\hat{f}(x)) = \frac{f(x)R(K)}{nh} + O\left(\frac{1}{n}\right).$$

En estas expresiones, $R(\phi) = \int_{\mathbb{R}} \phi(x)^2 dx$, $\sigma_K^2 = \int_{\mathbb{R}} u^2 K(u) du$. *Así, el error cuadrático medio es*

$$MSE(\hat{f}(x)) = \frac{f(x)R(K)}{nh} + \frac{(f''(x))^2 \sigma_K^4 h^4}{4} + O\left(\frac{1}{n}\right) + O(h^5).$$

Por lo tanto $\hat{f}(x) \longrightarrow f(x)$ *en probabilidad.*

Demostración:

$$
\begin{aligned}
E(\hat{f}(x)) &= E\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right) = E(K_h(x-X_i)) \\
&= \int_{\mathbb{R}} K_h(x-u)f(u)du = (K_h * f)(x) = \int_{\mathbb{R}}\frac{1}{h}K\left(\frac{x-u}{h}\right)f(u)du \\
&\quad \text{(por simetría de } K\text{)} \\
&= \int_{\mathbb{R}}\frac{1}{h}K\left(\frac{u-x}{h}\right)f(u)du \\
&\quad \text{(cambio de variable: } v=(u-x)/h,\ dv=(1/h)du\text{)} \\
&= \int_{\mathbb{R}} K(v)f(x+hv)dv \\
&\quad \text{(Taylor: } f(x+hv)=f(x)+f'(x)hv+\frac{1}{2}f''(x)h^2v^2+O(h^3)\text{)} \\
&= \int_{\mathbb{R}} K(v)(f(x)+f'(x)hv+\frac{1}{2}f''(x)h^2v^2)dv + O(h^3) \\
&= f(x)\int_{\mathbb{R}}K(v)dv + f'(x)h\int_{\mathbb{R}}vK(v)dv + \frac{1}{2}f''(x)h^2\int_{\mathbb{R}}v^2K(v)dv + O(h^3) \\
&= f(x) + \frac{f''(x)\sigma_K^2h^2}{2} + O(h^3).
\end{aligned}
$$

$$
\begin{aligned}
V(\hat{f}(x)) &= V\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right) = \frac{1}{n}V(K_h(x-X_i)) \\
&= \frac{1}{n}\left[E(K_h^2(x-X_i)) - E(K_h(x-X_i))^2\right] \\
&= \frac{1}{n}\left[\int_{\mathbb{R}} K_u^2(x-u)f(u)du - \left(\int_{\mathbb{R}} K_u(x-u)f(u)du\right)^2\right] \\
&= \frac{1}{n}\left[(K_h^2 * f)(x) - (K_h * f)^2(x)\right] \\
&\quad \text{(por simetría de } K\text{)} \\
&= \frac{1}{n}\int_{\mathbb{R}} \frac{1}{h^2}K^2\left(\frac{u-x}{h}\right)f(u)du - \frac{1}{n}\left(f(x)+O(h^2)\right)^2 \\
&\quad \text{(cambio de variable: } v=(u-x)/h,\ dv=(1/h)du\text{)} \\
&\quad \text{(él último sumando es } O(1/n)\text{)} \\
&= \frac{1}{nh}\int_{\mathbb{R}} K^2(v)f(x+hv)dv + O\left(\frac{1}{n}\right) \\
&\quad \text{(Taylor: } f(x+hv) = f(x)+O(h)\text{)} \\
&= \frac{1}{nh}\int_{\mathbb{R}} K^2(v)f(x)dv + \frac{1}{nh}O(h) + O\left(\frac{1}{n}\right) \\
&= \frac{f(x)}{nh}\int_{\mathbb{R}} K^2(v)dv + O\left(\frac{1}{n}\right) = \frac{f(x)R(K)}{nh} + O\left(\frac{1}{n}\right).
\end{aligned}
$$

□

Comportamiento global: MISE y AMISE

A partir de los resultados anteriores sobre el comportamiento local del estimador núcleo de la densidad, integrando sobre toda la recta real se obtiene lo siguiente:

$$\mathrm{MISE}(\hat{f}) = \int_{\mathbb{R}} \mathrm{MSE}(\hat{f}(x))dx = \frac{R(K)}{nh} + \frac{\sigma_K^4 h^4}{4}R(f'') + O\left(\frac{1}{n}\right) + O(h^5),$$

$$\mathrm{AMISE}(\hat{f}) = \frac{R(K)}{nh} + \frac{\sigma_K^4 h^4}{4}R(f'').$$

Derivando en h e igualando a 0, se obtiene que la ventana asintótica óptima (que minimiza el AMISE) es

$$h_0 = \left(\frac{R(K)}{\sigma_K^4 R(f'')}\right)^{\frac{1}{5}} n^{-\frac{1}{5}}.$$

El AMISE para la ventana óptima es

$$\text{AMISE}_0 = \frac{5}{4}\left(\sigma_K R(K)\right)^{\frac{4}{5}} R(f'')^{\frac{1}{5}} n^{-\frac{4}{5}}.$$

Observar que el AMISE óptimo es del mismo orden que en el caso del polígono de frecuencias, $O(n^{-4/5})$, mientras que el AMISE paramétrico es $O(n^{-1}) = o(n^{-4/5})$.

Ejemplo 3.7
Figura 2.5 de Wand y Jones (1995).

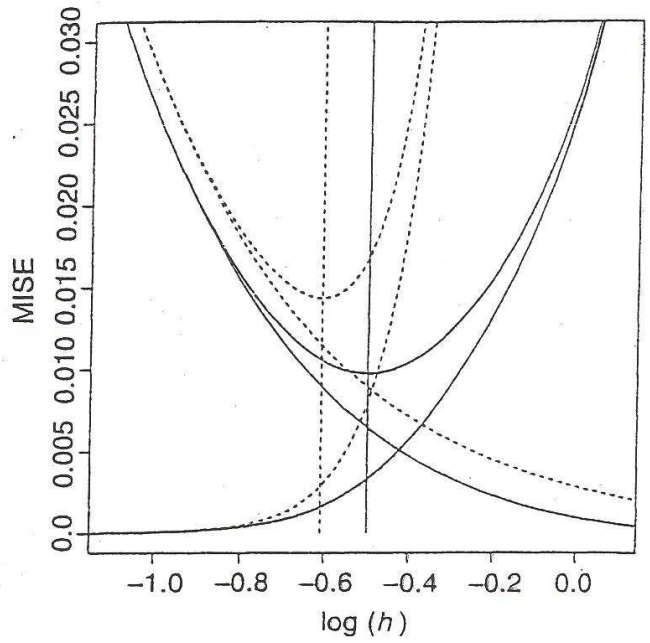


Figure 2.5. *Plots of* $\text{MISE}\{\hat{f}(\cdot;h)\}$ *(bowl-shaped solid curve) a*
$\text{AMISE}\{\hat{f}(\cdot;h)\}$ *(bowl-shaped dashed curve) versus* $\log_{10} h$ *f*
the density f_1 *and* $n = 100$. *Vertical lines are drawn throu*
their respective minimisers. Also plotted are the integrat
variance and its asymptotic approximation (decreasing solid a
dashed curves respectively) and the integrated squared bias a
its asymptotic approximation (increasing solid and dashed curv
respectively).

Eficiencia relativa de distintas funciones núcleo

Analicemos la expresión del AMISE óptimo:

$$\text{AMISE}_0 = \frac{5}{4}\left(\sigma_K R(K)\right)^{\frac{4}{5}} R(f'')^{\frac{1}{5}} n^{-\frac{4}{5}}.$$

El factor $R(f'')^{1/5}$ es una medida de la curvatura total de la función $f(x)$ que estamos estimando. Cuanto más curvatura tiene $f(x)$ mayor es el AMISE_0.

Por otra parte, el factor $\left(\sigma_K R(K)\right)^{\frac{4}{5}}$ sólo depende del núcleo K empleado en la construcción del estimador núcleo. Dado que tenemos libertad para elegir la función núcleo K, surge la pregunta de qué núcleo K hace menor esa cantidad. Eso equivale a plantearse el siguiente problema de cálculo de variaciones:

$$\begin{aligned} \min_K \quad & \sigma_K R(K) \\ s.a. \quad & \int_{\mathbb{R}} K(u)du = 1 \\ & \int_{\mathbb{R}} uK(u)du = 0 \\ & \int_{\mathbb{R}} u^2 K(u)du = a^2 \end{aligned}$$

En la tercera restricción se asigna un valor $a^2 < \infty$ arbitrario. Se puede comprobar que si no se fija un valor de este segundo momento el problema no está bien determinado porque se puede obtener el mismo valor de la función objetivo con los núcleos $K(u)$ y

$$K_\tau(u) = \frac{1}{\tau} K\left(\frac{u}{\tau}\right),$$

que sólo difieren en el parámetro de escala.

La solución del problema anterior para $a^2 = 1/5$ es el NÚCLEO DE EPANECHNIKOV:

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

El valor de la función objetivo para este núcleo es

$$\sigma_{K^*} R(K^*) = \frac{3}{5\sqrt{5}} = 0{,}2683.$$

La Figura 3.10 muestra la forma de algunas funciones núcleo usadas en estimación no paramétrica de la densidad. El Cuadro 3.1 muestra que la

Núcleo K	Fórmula	Eficiencia= $\sigma_{K^*}R(K^*)/\sigma_K R(K)$
Epanechnikov (K^*)	$(3/4)(1-x^2)I_{[-1,1]}(x)$	1
Biweight	$(15/16)(1-x^2)^2 I_{[-1,1]}(x)$	0.994
Triweight	$(35/32)(1-x^2)^3 I_{[-1,1]}(x)$	0.987
Gaussiano	$(1/\sqrt{2\pi})\exp(-x^2/2)$	0.951
Triangular	$(1-\|x\|)I_{[-1,1]}(x)$	0.986
Uniforme	$(1/2)I_{[-1,1]}(x)$	0.930

Cuadro 3.1: Eficiencia relativa de algunas funciones núcleo.

pérdida en eficiencia es muy pequeña si se usa un núcleo distinto al óptimo (el de Epanechnikov). La elección del núcleo debe obedecer más a cuestiones computacionales (mejor si tiene soporte compacto y si su evaluación no es muy costosa) o de suavidad (por ejemplo, el núcleo de Epanechnikov no es derivable en ± 1, mientras que los núcleos Biweight o Triweight sí lo son). Por último, cabe señalar que es mucho más importante la elección del parámetro de suavizado que la elección del núcleo.

Elección de la ventana mediante la regla de referencia a la normal

Si se supone que la densidad $f(x)$ corresponde a la de una $N(\mu, \sigma^2)$ y se usa un núcleo K_G gaussiano, la fórmula de la ventana óptima da este valor:

$$h_{0,K_G} = \left(\frac{R(K_G)}{\sigma^4_{K_G} R(f'')}\right)^{\frac{1}{5}} n^{-\frac{1}{5}} = 1{,}059\sigma n^{-\frac{1}{5}}.$$

El valor de $\sigma = \sqrt{V(X)}$ se estima a partir de los datos como ya vimos en el caso del histograma.

Si se usa otro núcleo K distinto al gaussiano la ventana óptima sería

$$h_{0,K} = \left(\frac{R(K)}{\sigma^4_K R(f'')}\right)^{\frac{1}{5}} n^{-\frac{1}{5}}$$

y se tiene que el cociente de la ventanas óptimas no depende de la función de densidad desconocida:

$$\frac{h_{0,K}}{h_{0,K_G}} = \left(\frac{R(K)/\sigma^4_K}{R(K_G)/\sigma^4_{K_G}}\right) = c_K.$$

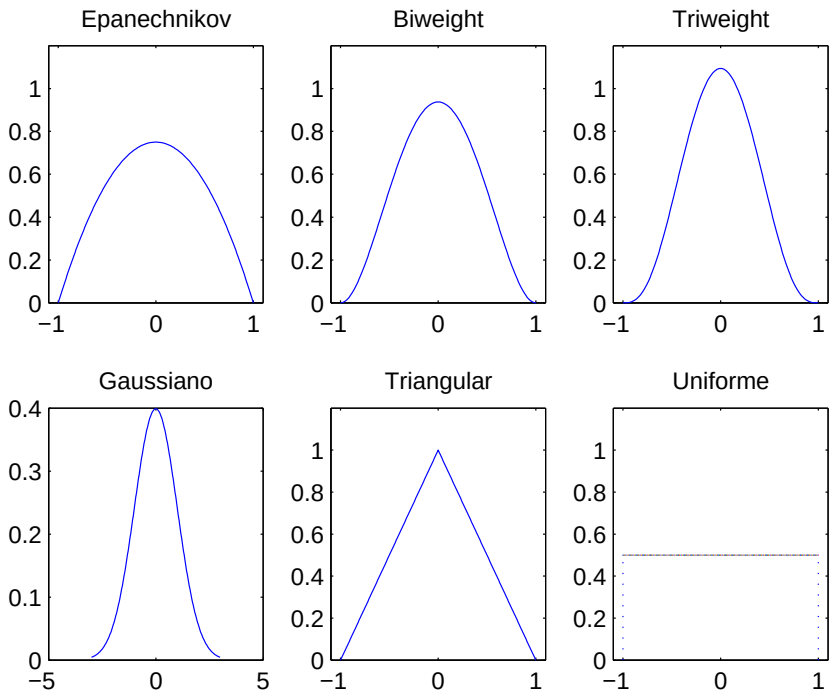


Figura 3.10: Algunos de los núcleos usados en estimación de la densidad.

Núcleo K	Factor c_K
Epanechnikov	2.214
Biweight	2.623
Triweight	2.978
Triangular	2.432
Uniforme	1.740

Cuadro 3.2: Constante c_K para algunas funciones núcleo.

Así,

$$h_{0,K} = c_K h_{0,K_G} = c_K 1{,}059 \sigma n^{-\frac{1}{5}}.$$

El Cuadro 3.2 recoge los valores de la constante c_K para distintas funciones núcleo.

Funciones núcleo con ventanas comparables

Algunos programas (por ejemplo la función `density` de R) utilizan versiones ajustadas (reescaladas) de las funciones núcleo usuales de tal modo que la varianza de todas estas versiones reescaladas sea igual a 1. De esta forma el parámetros de suavizado h (la ventana) tiene para todos las funciones núcleo el mismo significado: es la desviación típica del núcleo que se utiliza en el suavizado. Es decir, si $K(u)$ es un núcleo ya reescalado para tener varianza 1,

$$\sigma_K^2 = \int_{\mathbb{R}} u^2 K(u) du = 1$$

y $K_h(u) = (1/h)K(u/h)$ se tiene que (con el cambio de variable $v = u/h$)

$$\sigma_{K_h}^2 = \int_{\mathbb{R}} u^2 K_h(u) du = \int_{\mathbb{R}} u^2 \frac{1}{h} K\left(\frac{u}{h}\right) du = h^2 \int_{\mathbb{R}} v^2 K(v) dv = h^2.$$

Si $K_0(u)$ es un núcleo con varianza $\sigma_{K_0}^2$ el núcleo reescalado para que tenga varianza 1 es $K(u) = \sigma_{K_0} K_0(\sigma_{K_0} u)$.

El Cuadro 3.3 muestra las ecuaciones de los núcleos usuales reescaladas para que el parámetro h sea en todos ellos la desviación típica.

Las siguientes instrucciones dibujan en R las gráficas de estos núcleos. El resultado se muestra en la figura 3.11 (**Nota:** El núcleo *triweight* no está implementado; en su lugar se muestra el núcleo *cosine*.)

```
par(mfrow=c(3,2))
```

Núcleo	Fórmula original K	Varianza original $\sigma_K^2 = \int_{\mathbb{R}} u^2 K(u)du$	Fórmula reescalada
Epanechnikov	$(3/4)(1-x^2)I_{[-1,1]}(x)$	$1/5$	$(3/4\sqrt{5})(1-x^2/5)I_{[-\sqrt{5},\sqrt{5}]}(x)$
Biweight	$(15/16)(1-x^2)^2 I_{[-1,1]}(x)$	$1/7$	$(15/16\sqrt{7})(1-x^2/7)^2 I_{[-\sqrt{7},\sqrt{7}]}(x)$
Triweight	$(35/32)(1-x^2)^3 I_{[-1,1]}(x)$	$1/9$	$(35/96)(1-x^2/9)^3 I_{[-3,3]}(x)$
Gaussiano	$(1/\sqrt{2\pi})\exp(-x^2/2)$	1	$(1/\sqrt{2\pi})\exp(-x^2/2)$
Triangular	$(1-\lvert x\rvert)I_{[-1,1]}(x)$	$1/6$	$(1/\sqrt{6})(1-\lvert x\rvert/\sqrt{6})I_{[-\sqrt{6},\sqrt{6}]}(x)$
Uniforme	$(1/2)I_{[-1,1]}(x)$	$1/3$	$(1/2\sqrt{3})I_{[-\sqrt{3},\sqrt{3}]}(x)$

Cuadro 3.3: Ecuaciones de los núcleos usuales reescaladas.

Núcleo K	Factor c_K
Epanechnikov	1.01006
Biweight	1.00882
Triweight	0.99267
Triangular	1.00719
Uniforme	0.99540

Cuadro 3.4: Constante c_K para algunas funciones núcleo reescaladas.

```
nucleo <- c("epanechnikov", "biweight", "cosine",
            "gaussian", "triangular", "rectangular")
sapply(nucleo, function(a) plot(density(c(0),bw=1,kernel=a),main=a))
par(mfrow=c(1,1))
```

En cuanto a la elección óptima de la ventana, si se trabaja con núcleos ajustados para que tengan varianza 1 se tiene que

$$\frac{h_{0,K}}{h_{0,K_G}} = \frac{R(K)}{R(K_G)} = c_K \Rightarrow h_{0,K} = c_K h_{0,K_G} = c_K 1{,}059\sigma n^{-\frac{1}{5}}.$$

El cuadro 3.4 recoge los valores de la constante c_K para distintas funciones núcleo ajustadas.

3.3.2. Problemas de los estimadores núcleo y algunas soluciones

A continuación se enumeran algunos de los problemas prácticos que presentan los estimadores núcleo de la densidad. Muchos de estos problemas no son exclusivos de este tipo de estimadores.

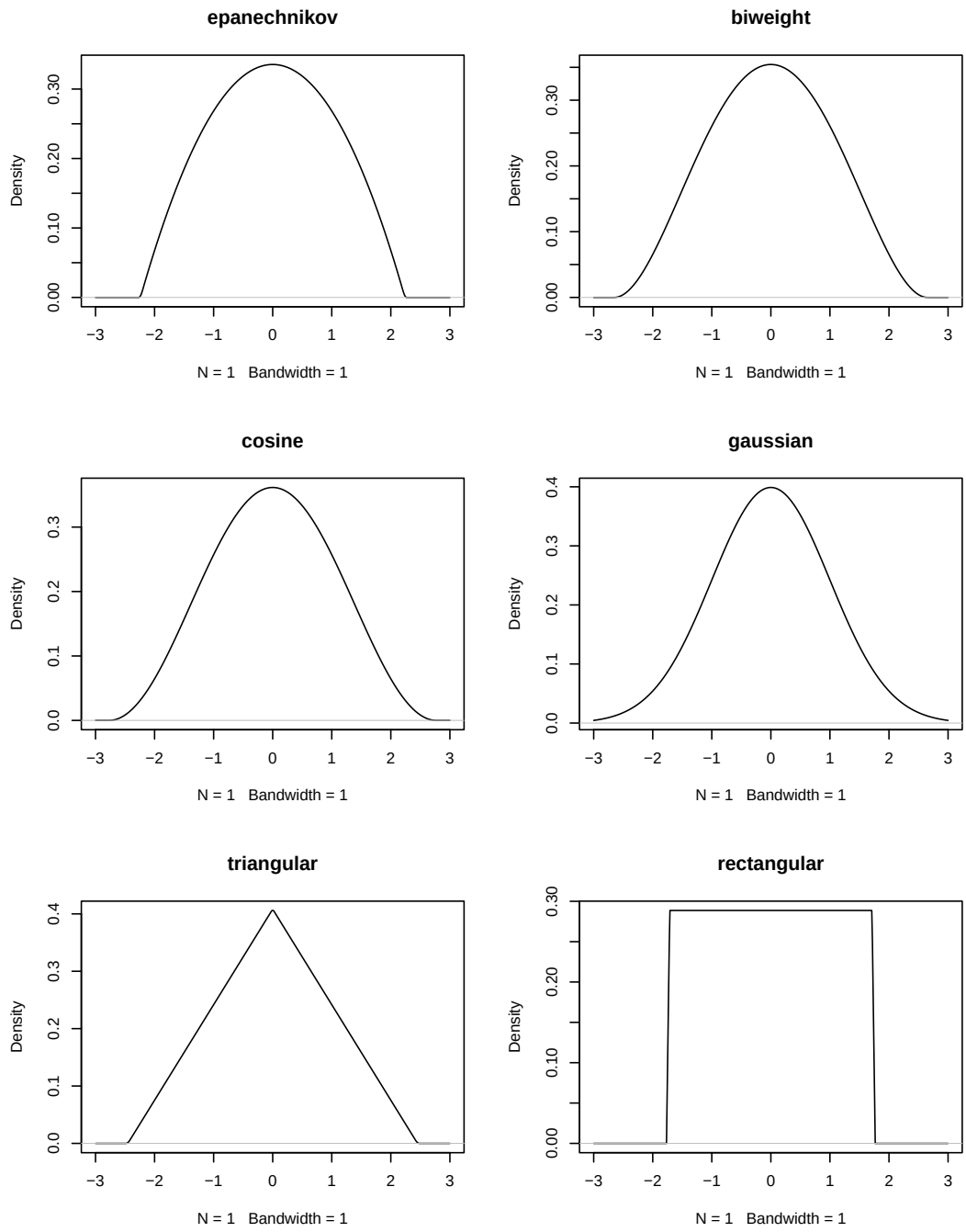


Figura 3.11: Gráficos en R de algunos de algunos núcleos reescalados.

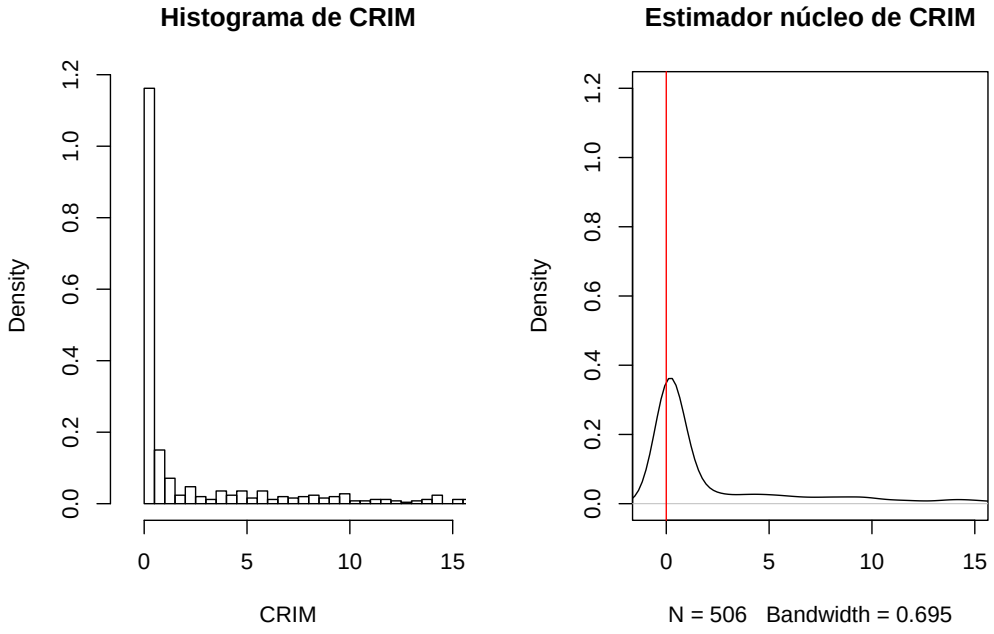


Figura 3.12: Problemas del estimador núcleo en el extremos del soporte de la densidad.

1. **Sesgo en los extremos del soporte de f, si éste es acotado.** La Figura 3.12 muestra la estimación núcleo de la densidad de la variable `CRIM`, tasa de criminalidad per cápita en los 506 barrios de Boston, junto con un histograma de la misma variable. A partir del histograma parece claro que la densidad debería ser decreciente en $[0, \infty]$, pero el estimador núcleo proporciona una densidad que no tiene su máximo en 0. Por otra parte, de la definición de la variable se sigue que su soporte es $[0, \infty]$, pero el estimador núcleo da probabilidad positiva a la semirrecta negativa.

   En la Figura 3.13 se representa la densidad de una exponencial de parámetro $\lambda = 1$ de la cual se ha extraído una muestra de tamaño $n = 20$. En el gráfico de la izquierda se muestra la función núcleo Gaussiana con $h = 0{,}2$ situada en torno al punto $x = 0{,}15$. Se aprecia cómo el estimador núcleo suaviza el máximo que la verdadera densidad tiene en el 0. Ello se debe a que el estimador núcleo compensa la alta densidad de puntos a la derecha del 0 con la densidad nula que hay a la izquierda de 0. Por otra parte, en el gráfico de la derecha se representa

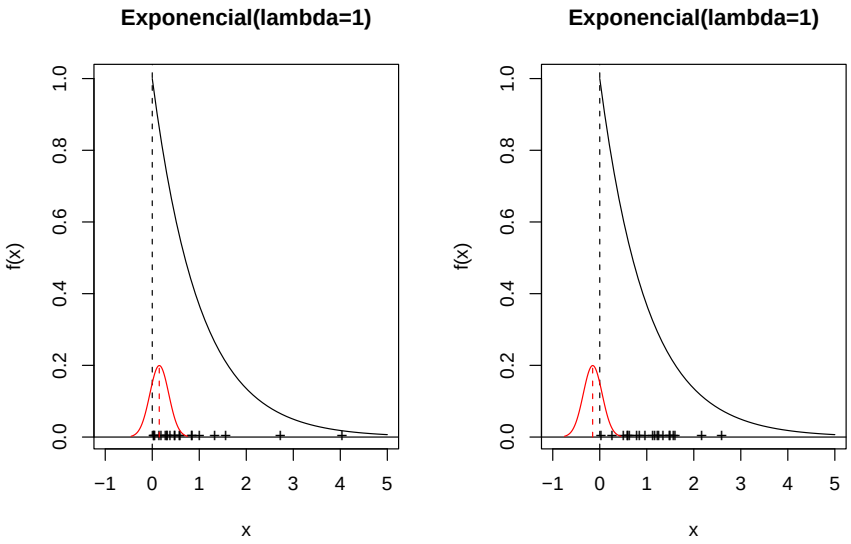


Figura 3.13: Estimación de la densidad cerca del 0 en una exponencial de parámetro $\lambda = 1$.

la función núcleo situada en torno al punto $x = -0{,}15$. Se observa cómo la estimación de la densidad es positiva en ese punto menor que 0.

Se puede probar que si X tiene soporte $[0, \infty)$ y K tiene soporte compacto en $[-1, 1]$, para $x \in [0, h)$ el estimador núcleo tiene sesgo de orden $O(1)$ (es decir, no va a 0 cuando n tiene a infinito), en lugar de ser de orden $O(h^2)$, como ocurre si $x \geq h$ o si el soporte de X es toda la recta real.

Una primera solución consiste en recortar y reponderar el núcleo cerca de la frontera del soporte de X para que dé peso 1 a ese soporte. Así, el estimador núcleo en un punto x cercano a la frontera (supongamos $x \in [0, h)$, como antes) sería

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{x,h}(x_i)$$

donde

$$K_{x,h}(x_i) = \frac{1}{\int_0^\infty \frac{1}{h} K\left(\frac{u-x}{h}\right) du} \frac{1}{h} K\left(\frac{x - x_i}{h}\right).$$

De este modo se consigue que el sesgo del estimador sea $O(h)$ en la frontera del soporte, una convergencia más lenta a 0 que el $O(h^2)$ del

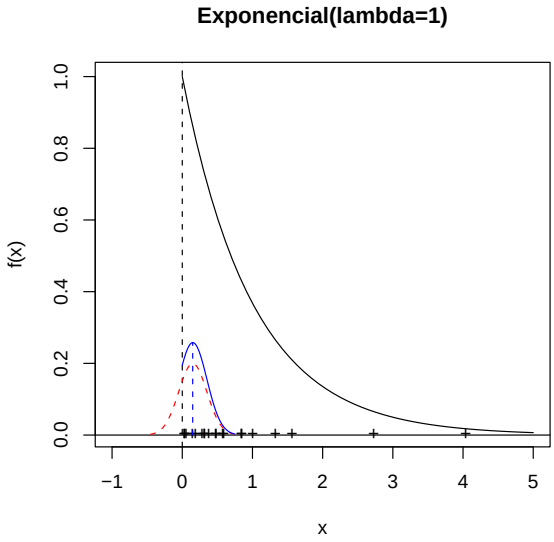


Figura 3.14: Estimación de la densidad cerca del 0.

sesgo en el interior del soporte. Por lo tanto recortar y renormalizar no es una solución totalmente satisfactoria. La Figura 3.14 ilustra este procedimiento en el caso de la estimación de la densidad de la exponencial de parámetro $\lambda = 1$.

2. **El estimador núcleo aplana picos y valles.**
   Recordemos la Figura 3.9. El gráfico de la derecha representa el valor esperado del estimador núcleo de la densidad representada en el panel de la izquierda. Se aprecia que en efecto el estimador núcleo aplana picos y valles. Esto se debe a que el sesgo del estimador núcleo es

$$\text{Sesgo}\hat{f}(x) = \frac{h^2\sigma_K^2 f''(x)}{2} + o(h^2).$$

   Por lo tanto el sesgo será positivo ($E(\hat{f}(x)) > f(x)$) en los puntos x donde $f''(x)$ sea positiva y grande en valor absoluto (allí donde haya un valle de la función de densidad), mientras que el sesgo será negativo en x cuando $f''(x)$ sea negativa y grande en valor absoluto (en los picos de la densidad).

3. **Falta de adaptación a las características locales de la función f.**
   El estimador núcleo usual no permite niveles diferentes de suavizado en

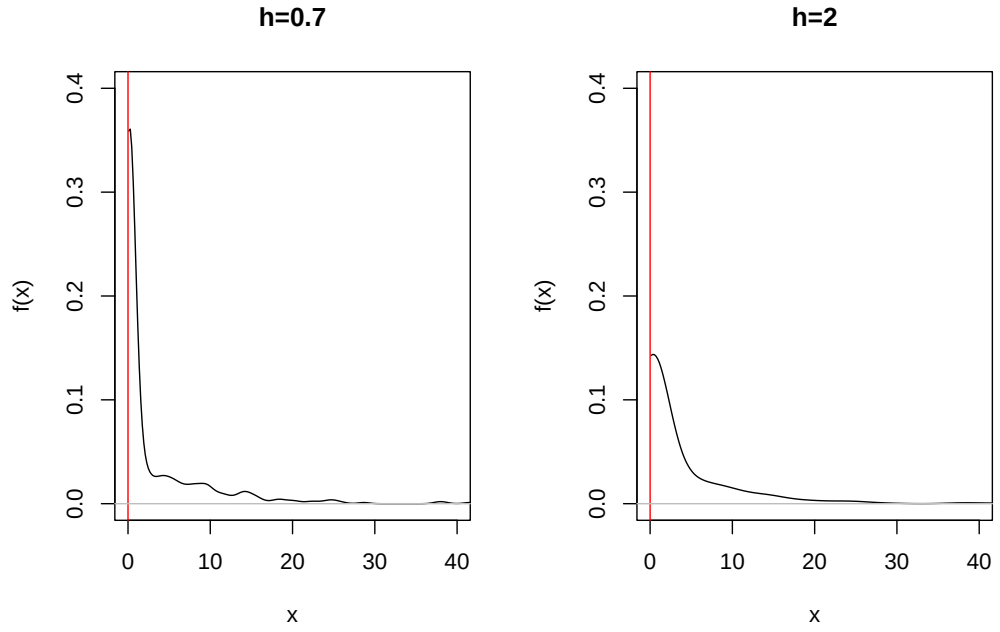


Figura 3.15: El estimador núcleo no permite adaptar el grado de suavizamiento a las características locales de $f(x)$.

partes diferentes del soporte de X. Por ejemplo, la Figura 3.15 muestra la estimación núcleo de la densidad de la variable `CRIM`, tasa de criminalidad per cápita en los 506 barrios de Boston, en el intervalo $[0, 40]$. Un valor pequeño de h hace un buen trabajo cerca del 0 (donde la densidad es alta), mientras que para valores grandes de x (donde la verdadera densidad es baja) la estimación es demasiado irregular. Un valor grande de h se adapta bien a las zonas donde la densidad es pequeña, pero no permite observar los detalles allí donde $f(x)$ es alta.

Una justificación de este fenómeno la ofrece la expresión del error cuadrático medio de $\hat{f}(x)$:

$$\text{MSE}(\hat{f}(x)) = \frac{f(x)R(K)}{nh} + \frac{(f''(x))^2\sigma_K^4 h^4}{4} + o\left(\frac{1}{nh}\right) + o(h^4).$$

Así que $\text{MSE}(\hat{f}(x))$ es creciente en $f(x)/h$ y en $h^4(f''(x))^2$.

Por lo tanto, para reducir el error cuadrático medio de $\hat{f}(x)$ el parámetro de suavizado h debería ser grande si $f(x)$ es grande y debería ser pequeño si $|f''(x)|$ es grande (zonas con mucha curvatura).

En la práctica, las desventajas de un estimador núcleo con h constante se traducen en un sobresuavizado de las zonas con mucha estructura ($|f''(x)|$ grande) y un infrasuavizado en las colas de la distribución (donde f es casi plana: $|f''(x)| \approx 0$).

A continuación se listan algunos de los posibles ajustes y modificaciones que pueden hacerse para corregir las deficiencias prácticas de los estimadores núcleo de la densidad que acabamos de enumerar.

1. **Funciones núcleo ajustadas en la frontera del soporte.**
   Se trata de encontrar funciones núcleo K_x específicas para la estimación de la densidad en cada punto x que diste menos de h de la frontera del soporte. Estos núcleos se deben ajustar de manera que el sesgo en la estimación de $f(x)$ sea $O(h^2)$, como ocurre en el resto del soporte de f. Para ello es necesario que el núcleo K_x tome valores negativos. Si bien es cierto que este método consigue reducir el sesgo en la frontera, esta reducción se hace a costa de aumentar la varianza de la estimación de $f(x)$ en esa zona. Para más detalles véase la Sección 3.3.1 de Simonoff (1996).

2. **Núcleos de alto orden.**
   Son núcleos K con momento de orden 2 nulo ($\int_{\mathbb{R}} u^2 K(u)du = 0$). Por lo tanto deben tomar valores negativos en algunas partes de su soporte. El hecho de tener ese segundo momento nulo hace que en el desarrollo de Taylor que permite aproximar el sesgo del estimador núcleo se anule el término de orden $O(h^2)$ y como resultado se tenga que el sesgo de un núcleo de alto orden sea de orden $O(h^4)$ o inferior. El efecto práctico principal es que el estimador se ajusta mejor a los picos y valles de la verdadera función de densidad.

3. **Estimador núcleo con ventana variable.**
   Como hemos visto antes, sería conveniente poder suavizar más donde $|f''(x)|$ sea pequeño, y menos donde ese valor sea grande. Esto se puede hacer, al menos, de dos formas:

   a) Estimador núcleo local. Se permite que la ventana h dependa del punto x donde se realiza la estimación:

   $$\hat{f}_L(x) = \frac{1}{nh(x)} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h(x)}\right).$$

   Obsérvese que el estimador $\hat{f}_L$ no es una verdadera función de densidad (no tiene por qué integrar 1).

Se puede probar que el valor $h(x)$ que minimiza el error cuadrático medio asintótico del estimador es

$$h_{\mathrm{AMSE}}(x) = \left(\frac{R(K) f(x)}{\sigma_K^4 f''(x)^2} \right)^{1/5} n^{-1/5}.$$

Por lo tanto, se necesita un estimador piloto de $f(x)$ para poder calcular en la práctica ese valor.

Un caso particular de estimador núcleo local es el conocido como ESTIMADOR NÚCLEO DE LOS k VECINOS MÁS CERCANOS (en inglés, *k-nearest neighbors kernel estimator*). En este estimador se usa un núcleo K con soporte en $[-1, 1]$ y la ventana $h(x) = h_k(x)$ se elige como el mínimo valor que permite que en $[x-h(x), x+h(x)]$ entren k de las observaciones x_i (serán las k observaciones más cercanas al punto x). El valor de k se puede elegir por *validación cruzada* (ver Sección 3.4). En ocasiones se fija la proporción de datos k/n que se quiere usar en la estimación de cada valor $f(x)$ en vez de fijar k. A esa proporción se le denomina *span* en algunos paquetes estadísticos. Ver la Sección 3.7 para más detalles sobre el estimador de los k vecinos más cercanos.

b) ESTIMADOR NÚCLEO DE VENTANA VARIABLE. Cada una de las observaciones x_i tiene asociado un valor de la ventana:

$$\hat{f}_V(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(x_i)} K\left(\frac{x - x_i}{h(x_i)} \right).$$

Obsérvese que $\hat{f}_V$ es una verdadera función de densidad.

Una buena elección de $h(x_i)$ es

$$h(x_i) = h_V f(x_i)^{-1/2}.$$

Así se reduce el sesgo de $O(h^2)$ a $O(h^4)$ y la varianza sigue siendo $O((nh)^{-1})$. Si además se toma $h_V = O(n^{-1/9})$ entonces MISE $= O(n^{-8/9})$.

Se necesita una estimación piloto de $f(x_i)$ para definir $h(x_i)$. Esta estimación piloto se realiza con una ventana fija. También hay que elegir el parámetro de suavizado h_V (ver Sección 3.4). En la práctica la elección de h_V tiene más influencia en el estimador final que la estimación piloto de $f(x_i)$.

4. **Estimación basada en una transformación.**
Supongamos que $x_1, \ldots, x_n$ es una muestra aleatoria simple de $X \sim f_X$,

y que la estimación de la función de densidad f_X presenta alguna dificultad (por ejemplo, su soporte es acotado, o tiene una moda muy pronunciada y al mismo tiempo colas altas). Es posible que transformando los datos obtengamos una densidad más fácilmente estimable (por ejemplo, el soporte de la variable aleatoria transformada puede ser toda la recta real).

Sea $Y = g(X)$, con g una transformación biyectiva del soporte de X en el soporte de Y. Entonces

$$f_Y(y) = f_X(g^{-1}(y))\frac{1}{|g'(g^{-1}(y))|},$$
$$f_X(x) = f_Y(g(x))|g'(x)|.$$

Los datos $y_i = g(x_i)$, $i = 1, \ldots, n$, forman una muestra aleatoria simple de $Y \sim f_Y$, que podemos usar para estimar no paramétricamente f_Y:

$$\hat{f}_Y(y) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right).$$

A partir de él, se define el ESTIMADOR NÚCLEO DE f_X BASADO EN LA TRANSFORMACIÓN g como

$$\hat{f}_{X,g}(x) = \hat{f}_Y(g(x))|g'(x)| = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{g(x) - g(x_i)}{h}\right)|g'(x)|.$$

Observar que, por el Teorema del valor Medio,

$$g(x) - g(x_i) = (x - x_i)g'(\psi_i)$$

para algún punto ψ_i intermedio entre x y x_i. Así,

$$\hat{f}_{X,g}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h/|g'(x)|} K\left(\frac{x - x_i}{h/|g'(\psi_i)|}\right).$$

Se ha supuesto que K es simétrico, para poder escribir $g'(\psi_i)$ en valor absoluto.

Se tiene entonces que $\hat{f}_{X,g}(x)$ es un híbrido entre $\hat{f}_L(x)$ y $\hat{f}_V(x)$ porque una ventana depende de x y otra de x y x_i conjuntamente.

Una familia de transformaciones que resulta útil es la de Box-Cox:

$$g(x; \lambda_1, \lambda_2) = \begin{cases} (x + \lambda_1)^{\lambda_2}\text{signo}(\lambda_2) & \text{si} \quad \lambda_2 \neq 0 \\ \log(x + \lambda_1) & \text{si} \quad \lambda_2 = 0 \end{cases}$$

con $\lambda_1 > -\inf\{\text{Soporte}(X)\}$. Se recomienda elegir λ_1 y λ_2 de manera que f_Y sea fácil de estimar.

3.4. Selección automática del parámetro de suavizado

Nos centraremos en la selección del parámetro de suavizado en el estimador núcleo, pero las ideas y métodos que se presentan se trasladan a otros estimadores con facilidad.

Esta sección se basa principalmente en el Capítulo 3 de Wand y Jones (1995). También puede consultarse la Sección 2.4 de Bowman y Azzalini (1997).

3.4.1. Regla de referencia a la normal

Este método ya ha sido presentado en la Sección 3.3, página 65. Recordemos únicamente que el valor del parámetro de suavizado h propuesto es

$$h_N = \left(\frac{8\sqrt{\pi}R(K)}{3\sigma_K^4}\right)^{1/5} \hat{\sigma} n^{-1/5} = c_K \cdot 1{,}059 \hat{\sigma} n^{-1/5},$$

con

$$\hat{\sigma} = \text{mín}\{S, \text{IQR}/1{,}35\},$$

donde S^2 es la varianza muestral de los datos, y IQR es su rango intercuartílico. La constante c_K depende del núcleo elegido y sus valores están recogidos en las tablas 3.2 y 3.4.

3.4.2. Sobresuavizado

Este método, del mismo modo que la regla de referencia a la normal descrita anteriormente, ofrece una fórmula fácil de evaluar y que en muchos casos proporciona valores razonables del parámetro de suavizado h. A veces se usan uno u otro método como punto de partida de reglas más complejas.

Recordemos que la expresión de h que minimiza el AMISE es

$$h_{\text{AMISE}} = \left(\frac{R(K)}{\sigma_K^4 R(f'')}\right)^{1/5} n^{-1/5}.$$

Si pudiésemos encontrar una cota inferior de $R(f'') = \int_{\mathbb{R}} (f''(x))^2 dx$, digamos R^*, entonces podríamos afirmar que

$$h_{\text{AMISE}} \leq \left(\frac{R(K)}{\sigma_K^4 R^*}\right)^{1/5} n^{-1/5} = h_{OS},$$

y el valor h_{OS} (OS viene de *oversmoothing*, que en inglés significa sobresuavizado) definido arriba será una cota superior de la ventana óptima h_{AMISE}. Por lo tanto, usando h_{OS} estamos seguros de que estamos sobresuavizando.

Scott (1992) se plantea el problema variacional siguiente:

$$\begin{aligned} \min_f \quad & R(f'') = \int_{\mathbb{R}} (f''(x))^2 dx \\ s.a. \quad & \int_{\mathbb{R}} f(x)dx = 1 \\ & \int_{\mathbb{R}} xf(x)dx = 0 \\ & \int_{\mathbb{R}} x^2 f(x)dx = 1 \end{aligned}$$

La solución es

$$f^*(x) = \frac{35}{96}(1 - x^2/9)^3 I_{[-3,3]}(x),$$

que es la expresión del núcleo Triweight reescalado parta tener varianza 1 (ver Tabla 3.3). Para esta función

$$R^* = R((f^*)'') = \frac{35}{243}.$$

Si se resuelve el problema con la restricción $\int_{\mathbb{R}} x^2 f(x)dx = \sigma^2$ la solución óptima es f^* con un cambio de escala:

$$\frac{1}{\sigma} f^* \left(\frac{x}{\sigma}\right)$$

y

$$R^* = \frac{35}{243}\frac{1}{\sigma^5}.$$

Por lo tanto,

$$h_{\text{AMISE}} \leq h_{OS} = \left(\frac{243R(K)}{35\sigma_K^4}\right)^{1/5} \sigma n^{-1/5}.$$

El valor de σ se estima a partir de los datos como en el caso de la regla de referencia a la normal.

Observar que

$$h_{OS} = \left(\frac{243/35}{8\sqrt{\pi}/3}\right)^{1/5} h_N = 1{,}08 h_N.$$

Por lo tanto la regla de referencia a la normal proporciona valores de h muy próximos al h de sobresuavizado. Se concluye que la regla de la normal tiende a sobresuavizar.

En la práctica el valor h_{OS} sirve de guía para empezar la exploración del h adecuado. Pueden probarse los valores h_{OS}, $h_{OS}/2$, $h_{OS}/4$, etc., y elegir entre esos valores el que se considere más adecuado visualmente. Naturalmente este proceso no puede automatizarse.

3.4.3. Validación cruzada por mínimos cuadrados

El criterio que hemos utilizado para medir la bondad de un estimador $\hat{f}(\cdot, h)$ (estimador núcleo con ventana h) de la función de densidad f es el error cuadrático integrado medio (MISE):

$$\text{MISE}(\hat{f}(\cdot,h)) = E_{\mathbf{X}}\left(\int_{\mathbb{R}}(\hat{f}(x,h)-f(x))^2dx\right) =$$

$$E_{\mathbf{X}}\left(\int_{\mathbb{R}}\hat{f}(x,h)^2dx\right) + \int_{\mathbb{R}}f(x)^2dx - 2E_{\mathbf{X}}\left(\int_{\mathbb{R}}\hat{f}(x,h)f(x)dx\right).$$

En esta expresión, $\mathbf{X}$ representa una m.a.s. de X de tamaño n, y $E_{\mathbf{X}}(\psi(\mathbf{X}))$ indica que se toma la esperanza de $\psi(\mathbf{X})$ con respecto a la distribución conjunta de $\mathbf{X}$. Observar que el segundo sumando no depende de h. Queremos buscar el h que minimice el $\text{MISE}(\hat{f}(\cdot,h))$, o de forma equivalente, que minimice

$$\text{MISE}(\hat{f}(\cdot,h)) - \int_{\mathbb{R}}f(x)^2dx = E_{\mathbf{X}}\left(\int_{\mathbb{R}}\hat{f}(x,h)^2dx - 2\int_{\mathbb{R}}\hat{f}(x,h)f(x)dx\right) \tag{3.2}$$

como función de h. Esta expresión depende de la función de densidad f desconocida a través de la segunda integral y a través de la esperanza en $\mathbf{X}$, cuya densidad conjunta es f^n. Por lo tanto no es posible evaluar esa expresión, dado que f es desconocida. Lo que sí es posible es construir un estimador de esa cantidad, que dependerá de h, y buscar el valor de h que haga mínimo ese estimador. Para ello conviene observar lo siguiente:

$$E_{\mathbf{X}}\left(\int_{\mathbb{R}}\hat{f}(x,h)f(x)dx\right)$$

$$= \int_{\mathbb{R}^n}\int_{\mathbb{R}}\hat{f}(x,h)f(x)\prod_{i=1}^{n}f(x_i)dxdx_1\ldots dx_n = E_{\mathbf{X},X}\left(\hat{f}(X,h)\right),$$

donde $X \sim f$ y además $\mathbf{X}$ y X son independientes.

Una primera propuesta para estimar

$$\int_{\mathbb{R}} \hat{f}(x,h)f(x)dx = E_X\left(\hat{f}(X,h)\right)$$

es hacerlo mediante el estadístico

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}(X_i,h),$$

pero aquí la variable aleatoria X_i donde se evalúa $\hat{f}(\cdot,h)$ no es independiente de las observaciones que hemos usado para construir el estimador no paramétrico de f.

Una alternativa es tomar como estimador

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i,h),$$

donde

$$\hat{f}_{-i}(x,h) = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{x-X_j}{h}\right).$$

De este modo garantizamos independencia entre el argumento del estimador no paramétrico y los datos usados para construirlo.

Así, se define el estadístico

$$\text{LSCV}(h) = \int_{\mathbb{R}} \hat{f}(x,h)^2dx - 2\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i,h)$$

y se toma el valor h que lo minimiza:

$$h_{LSCV} = \arg\min_h \text{LSCV}(h).$$

El procedimiento se llama *validación cruzada* porque se valida la estimación no paramétrica evaluando (de forma cruzada) el estimador construido con unas observaciones que no intervienen en su construcción. Esta metodología también se conoce como *leave-one-out* (dejar uno fuera).

Este procedimiento fue uno de los primeros intentos de buscar el parámetro h de forma automática. Una de las desventajas que presenta es que la función LSCV(h) puede tener varios mínimos locales. Las ventanas elegidas según este procedimiento presentan mucha variabilidad.

3.4.4. Plug-in directo

Hoy en día el método de elección de h que da mejores resultados prácticos es el conocido como *plug-in*. Se basa en sustituir en la expresión del valor h que minimiza el AMISE,

$$h_{\text{AMISE}} = \left(\frac{R(K)}{\sigma_K^4 R(f'')} \right)^{1/5} n^{-1/5},$$

la cantidad desconocida $R(f'')$ por una estimación hecha a partir de los datos observados.

Se puede probar que $R(f'') = \Psi_4(f)$, donde $\Psi_j(f) = E(f^{(j)}(X))$, $X \sim f$ y $f^{(j)}$ es la derivada j-ésima de f.

Por lo tanto, un estimador razonable de $R(f'')$ es

$$\hat{\Psi}_4(f) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{L,g}^{(iv)}(x_i),$$

donde $\hat{f}_{L,g}$ es un estimador de la función de densidad f construido con núcleo L y ventana g.

Surge inmediatamente la pregunta de cómo ha de ser elegida la ventana g para estimar $\Psi_4(f)$ de forma óptima (en el sentido de que minimice el AMSE de $\hat{\Psi}_4(f)$ como estimador de $\Psi_4(f)$).

Se puede probar que si se usa como L el mismo núcleo K original y f es suficientemente suave, entonces

$$g_{\text{AMSE}} = \left(\frac{-2K^{(iv)}(0)}{\sigma_K^4 \Psi_6(f)} \right)^{1/7} n^{-1/7}.$$

Esta regla tiene el mismo defecto que la regla plug-in para elegir h: necesitamos una estimación de $\Psi_6(f)$ para poder estimar g_{AMSE}.

Si se estima $\Psi_6(f)$ con el mismo núcleo K se llega a que la ventana óptima de la estimación depende de Ψ_8, y así sucesivamente: la ventana óptima para estimar Ψ_r depende de Ψ_{r+2}.

La estrategia habitual es estimar una de las Ψ_{4+2l} mediante la regla de referencia a la normal (calculando su valor si f fuese la densidad de una $N(\mu, \sigma^2)$).

Después se van estimando con las ventanas g_{AMSE} los valores Ψ_{4+2j}, $j = l-1, l-2, \ldots, 0$.

Finalmente se toma como ventana de plug-in directo con l pasos, el valor

$$h_{DPI,l} = \left(\frac{R(K)}{\sigma_K^4 \hat{\Psi}_4} \right)^{1/5} n^{-1/5}.$$

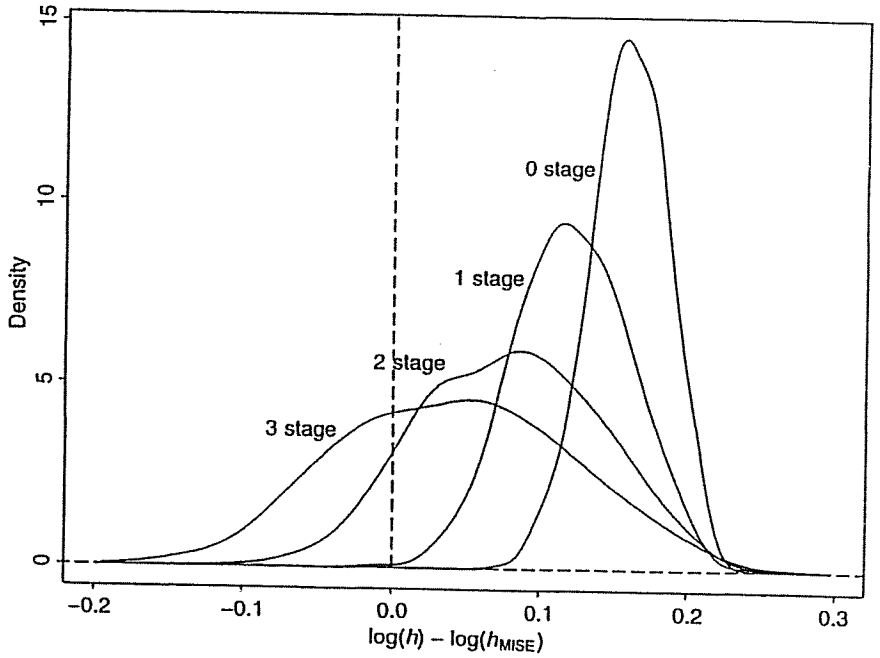


Figure 3.4. *Density estimates based on values of* $\log_{10}(\hat{h}_{\mathrm{DPI},\ell}) - \log_{10}(h_{\mathrm{MISE}})$ *for* $\ell = 0, 1, 2, 3$. *Selected bandwidths are based on 500 simulated samples of size* $n = 100$ *from the normal mixture density* f_1 *defined at (2.3).*

Figura 3.16: Comportamiento del selector de ventana $h_{DPI,l}$ para varios valores de l. (Figura 3.4 de Wand y Jones, 1995).

Observar que la regla de referencia a la normal que vimos anteriormente es también la regla de plug-in directo con $l = 0$ pasos:

$$h_N = h_{DPI,0}.$$

La Figura 3.16 muestra cómo la estimación de h mejora cuando l crece.

Es recomendable usar $h_{DPI,2}$. A este procedimiento de selección de la ventana también se le llama de Sheather y Jones, porque fueron estos autores quienes lo propusieron (Sheather y Jones 1991).

El siguiente algoritmo muestra cómo se implementaría este selector de ventana.

1. Estimar Ψ_8 suponiendo normalidad:

$$\hat{\Psi}_8^N = \frac{105}{32\sqrt{\pi}\hat{\sigma}^9}.$$

2. Estimar Ψ_6 mediante $\hat{\Psi}_6(g_1)$, donde

$$g_1 = \left(\frac{-2K^{(vi)}(0)}{\sigma_K^4 \hat{\Psi}_8^N(f)} \right)^{1/9} n^{-1/9}.$$

3. Estimar Ψ_4 mediante $\hat{\Psi}_4(g_2)$, donde

$$g_2 = \left(\frac{-2K^{(iv)}(0)}{\sigma_K^4 \hat{\Psi}_6(g_1)} \right)^{1/7} n^{-1/7}.$$

4. Seleccionar la ventana

$$h_{DPI,2} = \left(\frac{R(K)}{\sigma_K^4 \hat{\Psi}_4(g_2)} \right)^{1/5} n^{-1/5}.$$

3.4.5. Validación cruzada por máxima verosimilitud

Para la elección del parámetro de suavizado h existen versiones de los métodos conocidos como *validación cruzada*. En concreto, para un valor dado h se estima la verosimilitud de x_i a partir del estimador no paramétrico de la densidad calculado con el resto de la muestra y ese valor de h:

$$\hat{f}_{h,(-i)}(x_i) = \frac{1}{(n-1)h} \sum_{j \neq i}^{n} K\left(\frac{x_i - x_j}{h} \right).$$

Después se define la verosimilitud de la muestra por validación cruzada para el valor h del parámetro de suavizado como

$$L_{CV}(h) = \prod_{i=1}^{n} \hat{f}_{h,(-i)}(x_i)$$

y se toma como valor de h aquel h_{LCV} que hace máxima esa cantidad.

Este método no es el más utilizado, pese a la sencillez de su planteamiento. Véase la Sección 3.4.4 de Silverman (1986) para más detalles. En concreto se recomienda usarlo con cautela cuando puede haber datos atípicos.

3.4.6. Otros métodos

- Validación cruzada sesgada (*biased cross-validation*): h_{BCV}.
  Está relacionada con la validación cruzada por mínimos cuadrados. Se

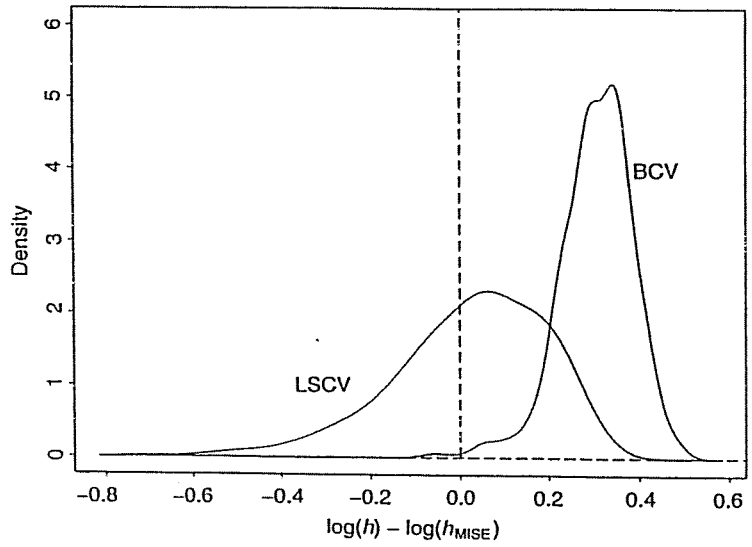


Figure 3.3. *Density estimates of* $\log_{10}(\hat{h}_{\mathrm{LSCV}}) - \log_{10}(h_{\mathrm{MISE}})$ *and* $\log_{10}(\hat{h}_{\mathrm{BCV}}) - \log_{10}(h_{\mathrm{MISE}})$. *Selected bandwidths are based on 500 simulated samples of size* $n = 100$ *from the normal mixture density* f_1 *defined at (2.3).*

Figura 3.17: Comparación del comportamiento de h_{LSCV} y h_{BCV} como estimadores de la ventana h_{MISE} que minimiza el MISE. (Figura 3.3 de Wand y Jones, 1995).

renuncia a estimar de forma insesgada la función objetivo (3.2) dependiente de h, pero se consigue estimarla con menor varianza.
Como consecuencia, h_{BCV} tiene menos varianza que h_{LSCV} (es más estable), aunque tiene algo de sesgo.
A veces el estimador de la función objetivo presenta más de un mínimo local.
En general, h_{BCV} es preferible a h_{LSCV}. Ver Figura 3.17.

- Regla que resuelve una ecuación (*solve-the-equation*): h_{STE}.
  Está relacionada con la regla plug-in directo. En la expresión que aproxima el valor de h que minimiza el AMISE,

$$\hat{h}_{\mathrm{AMISE}} = \left(\frac{R(K)}{\sigma_K^4 \hat{\Psi}_4(g)} \right)^{1/5} n^{-1/5},$$

se puede probar que el valor de g que estima bien Ψ_4 puede expresarse en términos de h_{AMISE}: $g = g(h_{\text{AMISE}})$.

Por lo tanto se propone resolver numéricamente en h la ecuación implícita siguiente:

$$h = \left(\frac{R(K)}{\sigma_K^4 \hat{\Psi}_4(g(h))} \right)^{1/5} n^{-1/5}.$$

Al valor resultante se le denota por h_{STE}.

Scott, Tapia y Thompson (1977) proponen resolver la ecuación implícita

$$h = \left(\frac{R(K)}{\sigma_K^4 \hat{\Psi}_4(h)} \right)^{1/5} n^{-1/5}.$$

Al valor resultante se le llama ventana de Scott, Tapia y Thompson (h_{STT}). Este método plantea problemas de convergencia.

- Validación cruzada suavizada (*smooth cross-validation*): h_{SCV}.
- Bootstrap suavizado: h_{SB}.

Comparaciones y recomendaciones prácticas

La Figura 3.18 compara distintos selectores automáticos del parámetro de suavizado en la estimación de la densidad de una mixtura de dos normales.

A partir de este ejemplo y de otros estudios publicados, se concluye que los selectores basados en plug-in directo y en resolver una ecuación son los más recomendables y que ambos son comparables entre sí. Tras ellos se sitúa el basado en validación cruzada sesgada y por último quedaría el basado en validación cruzada por mínimos cuadrados:

$$LSCV \prec BCV \prec \left\{ \begin{array}{l} DPI \\ STE \end{array} \right.$$

3.5. Estimación de la densidad multivariante

Sea ahora $x_1, \ldots, x_n$ una. m.a.s. de X, v.a. d-dimensional con densidad $f(x)$, $x \in \mathbb{R}^d$. Se desea estimar la función $f(x)$. Existen versiones del histograma y del polígono de frecuencias para estimar $f(x)$ con datos multivariantes, pero no son muy satisfactorias desde el punto de vista descriptivo. Por lo

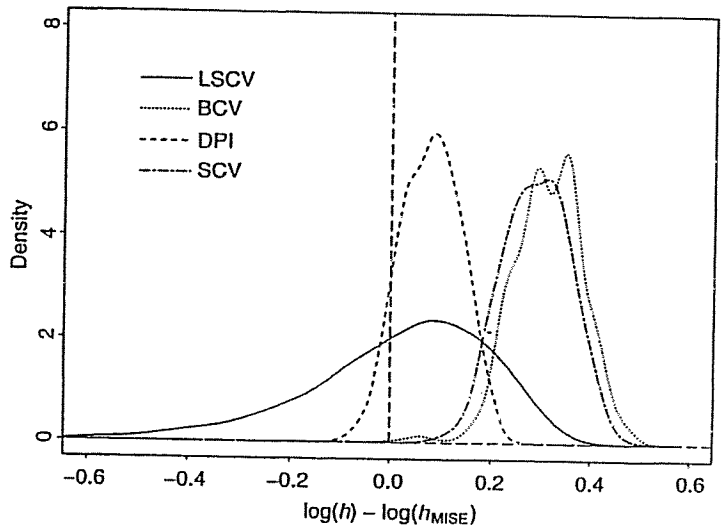


Figure 3.5. *Density estimates based on values of* $\log_{10}(\hat{h}) - \log_{10}(h_{\text{MISE}})$ *for several bandwidth selectors (described in the text). Selected bandwidths are based on 500 simulated samples of size* $n = 100$ *from the normal mixture density* f_1 *defined at (2.3).*

Figura 3.18: Comportamiento de varios selectores de ventana. (Figura 3.5 de Wand y Jones, 1995).

tanto presentaremos directamente el estimador núcleo de la densidad multivariante.

La generalización natural del estimador núcleo univariante es ésta,

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_d(x - x_i),$$

donde $K_d : \mathbb{R}^d \longrightarrow \mathbb{R}$ es una función núcleo d-dimensional que verifica:

$$\int_{\mathbb{R}^d} K_d(u)du = 1, \ \int_{\mathbb{R}^d} uK_d(u)du = 0 \in \mathbb{R}^d,$$

$$\int_{\mathbb{R}^d} uu^{\mathrm{T}} K_d(u)du = I_d \text{ (la matriz identidad } d \times d).$$

Usualmente se toma una función densidad centrada en el $0 \in \mathbb{R}^d$. Así, $\hat{f}(x)$ es una mixtura de n densidades, cada una de ellas centrada en una de las observaciones $x_i \in \mathbb{R}^d$.

Ejemplo 3.8

Consideremos el conjunto de datos referido a la vivienda en 506 barrios de Boston. La Figura 3.19 muestra el estimador núcleo bivariante de la densidad conjunta de las variables `LSTAT` y `RM` (número medio de habitaciones por domicilio) representada en 3 dimensiones y mediante curvas de nivel.

Se ha utilizado la librería `sm` en `R` y las instrucciones

```
sm.density(cbind(LSTAT,RM),h=c(1.5,.15),phi=30,theta=60,col=5)
sm.density(cbind(LSTAT,RM),h=c(1.5,.15),display="slice")
sm.density(cbind(LSTAT,RM),h=c(1.5,.15),display="slice",
           add=T,col=2,props=c(90))
```

.

La forma en la que se introduce distinto grado de suavizado en el estimador es mediante lo que se conoce como MATRIZ VENTANA H:

$$\hat{f}(x) = \frac{1}{n|H|}\sum_{i=1}^{n} K_d\left(H^{-1}(x - x_i)\right),$$

donde H es una matriz $d \times d$ no singular y $|H|$ es el valor absoluto del determinante de H. La matriz H representa una rotación de los datos y cambios de escala en cada variable. A continuación se listan algunos ejemplos de matrices H y los efectos que tiene su inclusión en el estimador:

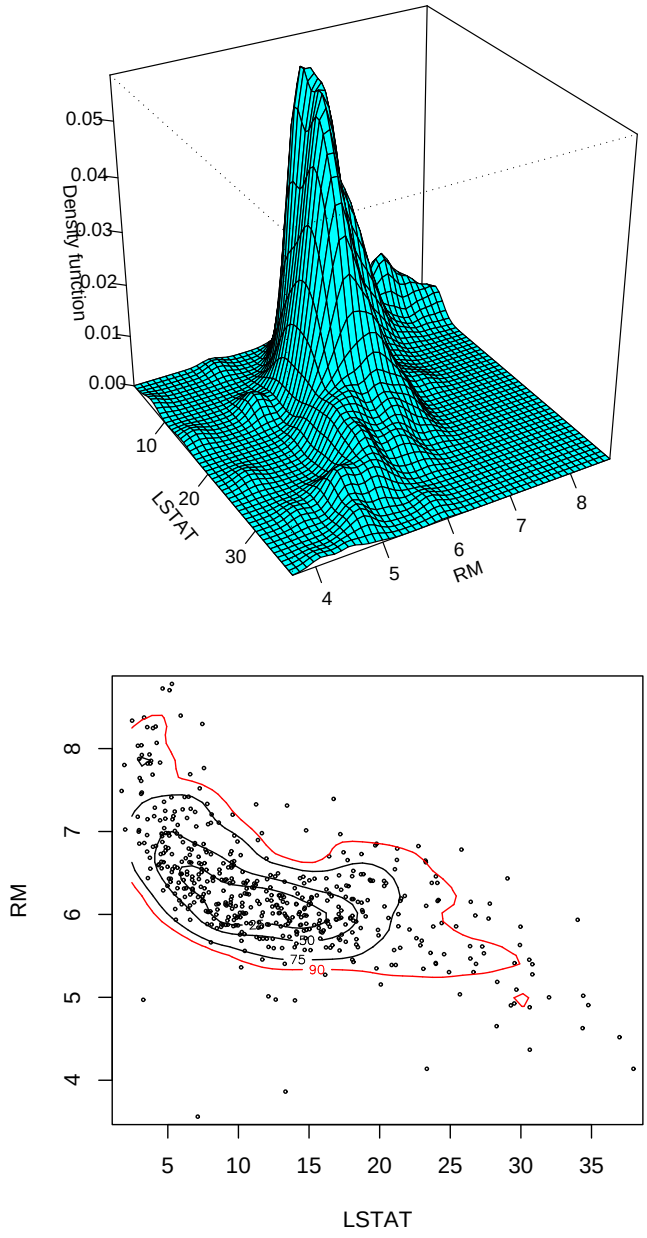


Figura 3.19: Estimador núcleo de la densidad conjunta de las variables `LSTAT` y `RM`.

- Un cambio de escala global: $H = hI_d$, $h > 0$. Se consigue mayor (h grande) o menor (h pequeño) grado de suavidad del estimador núcleo.
- Un cambio de escala distinto en cada dimensión: $H = \text{Diag}(h_1, \ldots, h_d)$, $h_j > 0$, $j = 1, \ldots, d$.
- Un cambio de escala distinto en cada dimensión seguido de una rotación: $H = \text{Diag}(h_1, \ldots, h_d)T$, $h_j > 0$, $j = 1, \ldots, d$, y Q una matriz ortonormal ($QQ^{\mathrm{T}} = Q^{\mathrm{T}}Q = I_d$). Se consigue diferente grado de suavidad en diferentes direcciones del espacio $\mathbb{R}^d$, que vienen dadas por las columnas de Q.

Una práctica usual es usar un núcleo producto que, dado K un núcleo univariante, se define como

$$K_d(u_1, \ldots, u_d) = \prod_{j=1}^{d} K(u_j).$$

Las propiedades asintóticas del estimador núcleo se resumen en la siguiente proposición.

Proposición 3.1 *Se escribe $H = hA$, con $|A| = 1$ y $h = |H|^{1/d} \in \mathbb{R}$. Entonces*

$$AMISE(\hat{f}) = \frac{R(K_d)}{nh^d} + \frac{h^4}{4} C(A, f),$$

donde $C(A, f)$ es una constante que depende sólo de f y de A. Si h tiende a 0 y nh^d tiende a infinito cuando n tiende a infinito, entonces $\hat{f}(x)$ converge puntualmente en MSE y globalmente en MISE.

El valor de h que minimiza el AMISE y el AMISE correspondiente son

$$h_0 = O(n^{-1/(d+4)}), \ AMISE_0 = O(n^{-4/(d+4)}).$$

La principal consecuencia de este resultado es que cuanto mayor es la dimensión d de los datos, menor es la precisión con la que se estima la densidad. Eso se debe a que cuanto mayor es d más alejado de -1 es el orden de convergencia del AMISE_0.

3.5.1. Elección de la matriz ventana

La matriz ventana H tiene d^2 elementos que hay que seleccionar. En la práctica las tres formas usuales de elegir esta matriz son las siguientes.

1. $H = hI_d$.
   Es razonable si la escala de las variables es comparable o si se ha estandarizado previamente. El valor de h se puede elegir por algún criterio análogo a los vistos en dimensión 1 (plug-in o validación cruzada, por ejemplo).

2. $H = \text{Diag}(h_1, \ldots, h_d)$.
   Es la opción más frecuente en los paquetes estadísticos que incorporan estimación no paramétrica de densidades multivariantes. Si además se usa un núcleo producto, el estimador núcleo de la densidad queda de la forma
   $$\hat{f}(x) = \frac{1}{n \prod_{j=1}^{d} h_j} \sum_{i=1}^{n} \prod_{j=1}^{d} K\left(\frac{x_j - x_{ij}}{h_j}\right)$$
   donde h_j es un parámetro de suavizado adecuado para la j-ésima coordenada de X.

   Si se toma núcleo producto gaussiano, el estimador de la densidad d dimensional será la mixtura de n densidades normales multivariantes con d coordenadas independientes con varianzas h_j^2, y cada densidad estará centrada en una observación x_i.

   La elección de cada h_j puede hacerse al menos de tres formas:

   a) Se considera la muestra univariante formada por la componente j-ésima de los datos observados y a partir de ellos se elige h_j con alguno de los métodos univariantes (plug-in, por ejemplo). Después se ajusta ese valor por el cambio de dimensión:
   $$h_j = h_j^{\text{unidim}} \frac{n^{-1/(d+4)}}{n^{-1/5}}.$$

   b) Se hace $h_j = a\hat{\sigma}_j$, $j = 1, \ldots, d$, y se utiliza algún criterio (plug-in, por ejemplo) análogo a los vistos en dimensión 1 para elegir el valor de a.

   c) Regla de la normal: $h_j = 1{,}05\hat{\sigma}_j n^{-1/(d+4)}$.

3. $H = hC^{1/2}$, donde C es la matriz de covarianzas de los datos.
   Esta es una forma de tener en cuenta la correlación entre las coordenadas de X. En vez de tomar núcleos multivariantes con coordenadas independientes (es lo que ocurre si tomamos el producto de núcleos univariantes) se toma como núcleo la función de densidad de una variable aleatoria cuya matriz de varianzas y covarianzas sea un múltiplo

h^2 de la matriz de varianzas y covarianzas muestral C de los datos $(x_{i1}, \ldots, x_{id})$, $i = 1, \ldots, n$.

Por ejemplo, si se toma un núcleo gaussiano multivariante con estas características se tiene que

$$\hat{f}_K(x) = \frac{1}{n(2\pi)^{d/2}h^d|C|^{1/2}} \sum_{i=1}^{n} \exp\left\{-\frac{1}{2h}(x - x_i)^{\mathrm{T}} C^{-1}(x - x_i)\right\}.$$

El valor de h se elige mediante alguno de los criterios conocidos.

Existe una versión multivariante de la regla de la normal:

$$H_N = \left(\frac{4}{d+2}\right)^{1/(d+4)} \Sigma^{1/2} n^{-1/(d+4)},\ \Sigma = \mathrm{Var}(X).$$

Dado que $(4/(d+2))^{1/(d+4)} \in (0{,}924,\ 1{,}059)$, ese valor se aproxima por 1. Así, en la práctica

$$H_N = C^{1/2} n^{-1/(d+4)}.$$

3.5.2. Representación de densidades tri-variantes

Sea X es una variable aleatoria de dimensión $d = 3$ con densidad $f(x)$, $x \in \mathbb{R}^3$. No es obvio cómo representar $f(x)$ gráficamente.

Una opción consiste en representar los CONTORNOS DE NIVEL, que son análogos a las curvas de nivel cuando $d = 2$:

$$C_k = \{x \in \mathbb{R}^3 : f(x) = k\} \subseteq \mathbb{R}^3.$$

Los conjuntos C_k son superficies bidimensionales inmersas en $\mathbb{R}^3$ y pueden ser representadas mediante técnicas estándar.

Ejemplo 3.9

Para el conjunto de datos referido a la vivienda en 506 barrios de Boston, la Figura 3.20 muestra el estimador núcleo trivariante de la densidad conjunta de las variables `CMEDV` (valor medio de las casas ocupadas por sus propietarios, en miles de dólares), `LSTAT` y `RM` mediante un contorno de nivel que encierra una probabilidad de 0.75.

Se ha utilizado la librería `sm` en `R` y la instrucción

```
sm.density(cbind(LSTAT,CMEDV,RM),h=c(1.5,1.5,.15),
           theta=-40,props=75)
```

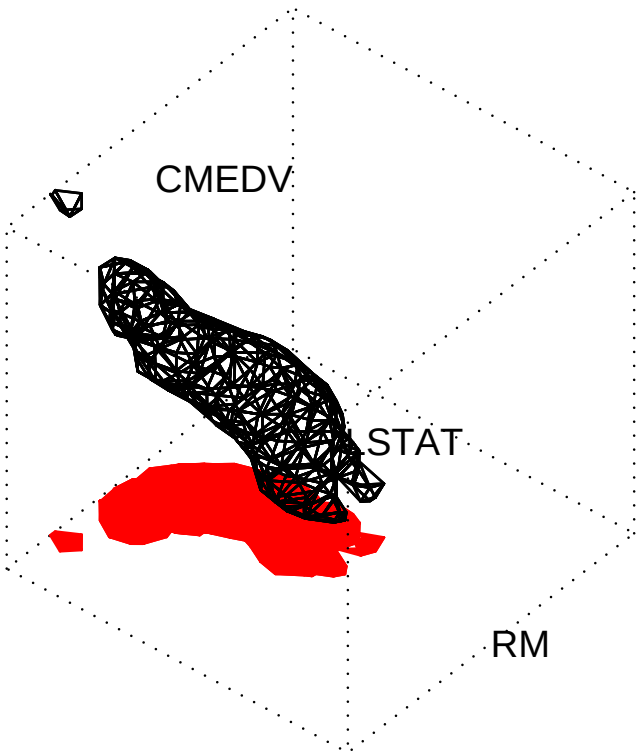


Figura 3.20: Contorno de nivel 0.75 de la densidad conjunta de las variables `CMEDV`, `LSTAT` y `RM`.

Otra opción es representar las densidades bivariantes de dos de las tres variables, condicionando a que la tercera variable pertenece a distintos intervalos. Este método recibe el nombre de GRÁFICOS CONDICIONALES. Es la única alternativa para dimensiones $d \geq 4$.

Ejemplo 3.10

Para el conjunto de datos referido a la vivienda en 506 barrios de Boston, la Figura 3.20 muestra la densidad conjunta de las variables `CMEDV`, `LSTAT` y `RM` mediante gráficos de la densidad de `CMEDV` y `LSTAT` condicionales a tres rangos de valores de la variable `RM`.

Se ha utilizado la librería `sm` en `R` y las instrucciones

```
   par(mfrow=c(2,2))
sm.density(cbind(LSTAT,CMEDV),h=c(1.5,1.5),display="slice")
title(main="All data")
q33 <- quantile(RM,.33)
q66 <- quantile(RM,.66)
I1<-(RM<q33)
sm.density(cbind(LSTAT[I1],CMEDV[I1]),h=c(1.5,1.5),
    display="slice",xlim=c(0,40),xlab="LSTAT",ylab="CMEDV")
title(main="RM<Q(.33)")
I2<-( (RM>=q33) & (RM<q66) )
sm.density(cbind(LSTAT[I2],CMEDV[I2]),h=c(1.5,1.5),
    display="slice",xlim=c(0,40),xlab="LSTAT",ylab="CMEDV")
title(main="Q(.33)<=RM<Q(.66)")
I3<-(RM>=q66)
sm.density(cbind(LSTAT[I3],CMEDV[I3]),h=c(1.5,1.5),
    display="slice",xlim=c(0,40),xlab="LSTAT",ylab="CMEDV")
title(main="Q(.66)<=RM")
```

3.5.3. La maldición de la dimensionalidad

Hemos visto más arriba que cuanto mayor es la dimensión d de los datos, menor es la precisión con la que se estima la función de densidad (el orden de convergencia del AMISE_0 se aleja de -1 cuando d crece).

Ésta es sólo una muestra de la dificultad que entraña hacer estimación no paramétrica de la densidad (y de otras funciones) cuando la dimensión d de los datos es grande.

En inglés este problema se conoce como *the curse of dimensionality*, que puede traducirse como LA MALDICIÓN DE LA DIMENSIONALIDAD. Se debe a que en dimensiones altas los entornos de un punto están prácticamente vacíos de los puntos observados en una muestra.

Dicho de otro modo: si queremos construir una bola centrada en un punto $x_0 \in \mathbb{R}^d$ que contenga digamos el 25 % de los puntos observados, esa bola deberá ser tan grande que difícilmente podremos decir que representa un entorno de x_0.

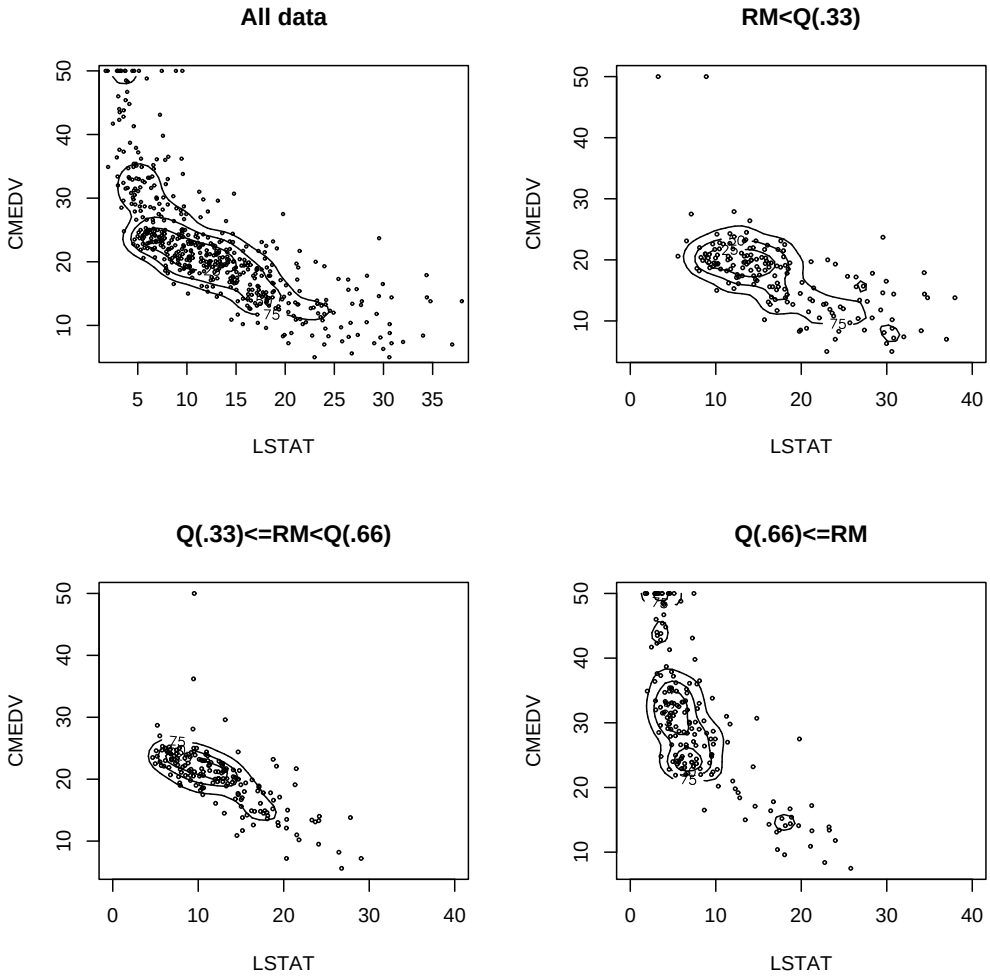


Figura 3.21: Representación de la densidad conjunta de las variables `CMEDV`, `LSTAT` y `RM` mediante gráficos de la densidad de `CMEDV` y `LSTAT` condicionales a 3 niveles de la variable `RM`. En el primer gráfico se muestra la densidad de `CMEDV` y `LSTAT` sin condicionar.

La siguiente tabla muestra este efecto. Sea $X \sim U([-1,1]^d)$ una variable aleatoria uniforme en el hiper-cubo d dimensional de lado $[-1,1]$. Sea $B_d(0_d,1)$ la bola centrada en el origen de $\mathbb{R}^d$ de radio 1. Para distintas dimensiones d se calcula la probabilidad que X pone en esa bola, que podríamos considerar un entorno de 0_d.

d	$P(X \in B_d(0_d,1))$
1	1
2	0.79
⋮	⋮
5	0.16
⋮	⋮
10	0.0025

3.6. Inferencia basada en la estimación de la densidad

En esta sección se tratan algunos problemas de inferencia estadística que pueden abordarse mediante la estimación no paramétrica de funciones de densidad. La referencia básica para esta sección es el Capítulo 2 de Bowman y Azzalini (1997) y la librería `sm` de `R` (que acompaña el citado texto).

3.6.1. Bandas de variabilidad

Recordemos que el sesgo y la varianza puntuales del estimador núcleo

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

son aproximadamente

$$\text{Sesgo}(\hat{f}(x)) \approx \frac{h^2\sigma_K^2}{2}f''(x), \;\; \text{Var}(\hat{f}(x)) \approx \frac{f(x)R(K)}{h}.$$

Se puede probar que, para h fijo, el estimador núcleo es asintóticamente normal:

$$\hat{f}(x) \sim AN\left(f(x) + \frac{h^2\sigma_K^2}{2}f''(x),\; \frac{f(x)R(K)}{nh}\right).$$

Si $f''(x)$ y $f(x)$ fuesen conocidos se conocería aproximadamente el sesgo y la varianza de $\hat{f}(x)$ y se podrían dar intervalos de confianza asintóticos

$(L(x), U(x))$ para $f(x)$ (aunque si $f(x)$ fuese conocido no tendría mucho interés dar intervalos de confianza para $f(x)$). En ese caso, diríamos que las funciones $\{(L(x), U(x)) : x \in \mathbb{R}\}$ son *bandas de confianza puntuales* para la función $f(x)$.

Dado que $f''(x)$ y $f(x)$ son desconocidos, no será posible construir dichas bandas de confianza, y nos limitaremos a dar *bandas de variabilidad puntuales* para $f(x)$ (que se definirán como bandas de confianza para $E(\hat{f}(x))$).

Nos basaremos en que

$$\hat{f}(x) \sim AN\left(E(\hat{f}(x)),\ \frac{f(x)R(K)}{nh}\right)$$

y en el Teorema δ de Cramer, también conocido como MÉTODO DELTA.

Teorema 3.4 (Método delta) *Si $X_n \sim AN(a, b_n)$ y $t : \mathbb{R} \longrightarrow \mathbb{R}$ es derivable dos veces con t'' continua en a, entonces*

$$t(X_n) \sim AN(t(a), (t'(a))^2 b_n).$$

La demostración se basa en un desarrollo de Taylor de primer orden: $t(X_n) \approx t(a) + t'(a)(X_n - a)$.

Así, si t es una función con dos derivadas continuas,

$$t(\hat{f}(x)) \sim AN\left(t(E(\hat{f}(x))),\ (t'(E(\hat{f}(x))))^2 \frac{f(x)R(K)}{nh}\right).$$

En la expresión de la varianza asintótica consideraremos despreciable el sesgo asintótico ($E(\hat{f}(x)) \approx f(x)$). Tenemos entonces que

$$t(\hat{f}(x)) \sim AN\left(t(E(\hat{f}(x))),\ (t'(f(x)))^2 \frac{f(x)R(K)}{nh}\right).$$

Para que la distribución asintótica (o al menos su varianza) no dependa de la función desconocida $f(x)$ deberíamos elegir la función t tal que $(t'(f(x)))^2 f(x)$ fuese constante en x.

Eso se consigue si tomamos $t(y) = \sqrt{y}$, que tiene $t'(y) = 1/(2\sqrt{y})$. En este caso la expresión anterior se particulariza así:

$$\sqrt{\hat{f}(x)} \sim AN\left(\sqrt{E(\hat{f}(x))},\ \frac{R(K)}{4nh}\right).$$

Así, un intervalo de confianza $(1-\alpha)$ asintótico para $\sqrt{E(\hat{f}(x))}$ será

$$\left(\sqrt{\hat{f}(x)} \mp z_{\alpha/2}\sqrt{\frac{R(K)}{4nh}}\right), x \in \mathbb{R},$$

donde z_p es el cuantil $(1-p)$ de una $N(0,1)$. Elevando al cuadrado ambos extremos obtenemos lo que llamaremos BANDAS DE VARIABILIDAD PUNTUALES para $f(x)$:

$$\left(\left(\sqrt{\hat{f}(x)} - z_{\alpha/2}\sqrt{\frac{R(K)}{4nh}}\right)^2, \left(\sqrt{\hat{f}(x)} + z_{\alpha/2}\sqrt{\frac{R(K)}{4nh}}\right)^2\right), x \in \mathbb{R}.$$

Estas bandas dan una idea de cómo es la variabilidad del estimador no paramétrico de la densidad f. Son bandas puntuales, no uniformes. Recordemos que no son bandas de confianza, sino bandas de variabilidad.

La función `sm.density` de la librería `sm` de `R` dibuja estas bandas de variabilidad en torno a la estimación $\hat{f}(x)$ si se usa la opción `display="se"`.

Ejemplo 3.11

Consideremos de nuevo el conjunto de datos referido a la vivienda en 506 barrios de Boston. La Figura 3.22 muestra las bandas de variabilidad para el estimador de la densidad de la variables `LSTAT` (porcentaje de población con estatus social en la categoría inferior).

.

3.6.2. Contraste de normalidad

Sea $x_1, \ldots, x_n$ una m.a.s. de $X \sim f(x)$. Se quiere hacer un contraste de bondad de ajuste de la distribución normal a los datos observados. Dicho de otro modo, se desea contrastar si X es normal:

$$\begin{cases} H_0 : X \sim N(\mu, \sigma^2) \text{ para algunos } \mu, \sigma^2 \text{ desconocidos,} \\ H_1 : X \not\sim N(\mu, \sigma^2) \text{ para ningunos } \mu, \sigma^2. \end{cases}$$

Hay muchas formas de contrastar normalidad: tests basados en los coeficientes de asimetría y curtosis, tests de bondad de ajuste de la χ^2 o de Kolmogorov-Smirnov (test de Lilliefors), test gráfico del QQ-plot, etc.

También es posible definir un contraste de normalidad basado en la estimación no paramétrica de la densidad. Las ventajas que presenta son éstas:

- Puede detectar desviaciones de la normalidad que tengan una interpretación más intuitiva (bimodalidad, asimetría, apuntamiento de la moda no acorde con la normalidad, por exceso o por defecto, etc.) que las detectadas por otros métodos. Ello se debe a que la estimación de la densidad permite visualizar esas características y compararlas con las de la normal.

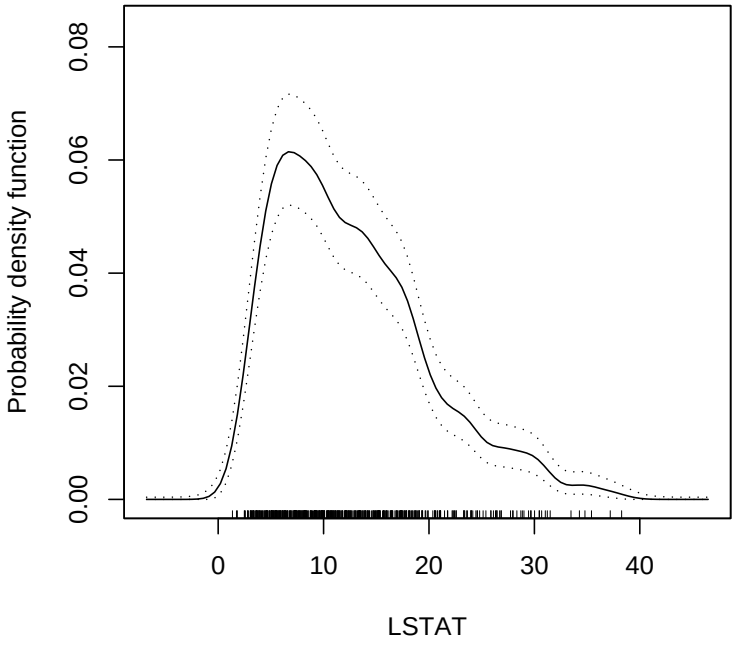


Figura 3.22: Estimador núcleo de la densidad de la variable `LSTAT`, acompañado de bandas de variabilidad.

- Puede generalizarse a dimensiones mayores que 1 más fácilmente que otros contrastes.

En general, un contraste de bondad de ajuste de un modelo paramétrico basado en la estimación no paramétrica de la densidad tiene esta estructura. Se trata de contrastar

$$H_0 : f \in \mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}, \text{ frente a } H_1 : f \notin \mathcal{F}_\Theta$$

Se usa como estadístico del contraste

$$T = d(f_{\hat{\theta}}, \hat{f}),$$

donde $\hat{\theta}$ es un estimador de θ (y por lo tanto $f_{\hat{\theta}}$ es un estimador paramétrico de f), $\hat{f}$ es un estimador no paramétrico de f y $d(\cdot, \cdot)$ es una distancia entre funciones de densidad.

La distribución de T bajo la hipótesis nula es desconocida y puede ser difícil de aproximar.

Veamos cómo puede hacerse esta aproximación en el caso de un contraste de normalidad.

Sea $f_N(x;\mu,\sigma^2)$ la función de densidad de una $N(\mu,\sigma^2)$. Recordemos que si $\hat{f}(x)$ es el estimador núcleo de f con núcleo K y ventana h, entonces su valor esperado es $(K_h * f)(x)$.

Bajo la hipótesis nula $f(x) = f_N(x;\mu,\sigma^2)$. Si se usa como K el núcleo Gaussiano con desviación típica h, la densidad $(K_h * f)(x)$ corresponde a la suma de una $N(\mu,\sigma^2)$ y de una $N(0,h^2)$ independientes, es decir,

$$(K_h * f)(x) = f_N(x;\mu,\sigma^2+h^2).$$

Si suponemos H_0 cierta, tomaremos h mediante la regla de la normal.

Así, es recomendable comparar $\hat{f}(x)$ con $f_N(x;\hat{\mu},\hat{\sigma}^2+h^2)$, en vez de hacer la comparación directa entre $\hat{f}(x)$ y $f_N(x;\hat{\mu},\hat{\sigma}^2)$.

Como distancia se puede usar la norma L_2. En ese caso el estadístico del test será

$$T = \int_{\mathbb{R}} \left(\hat{f}(x) - f_N(x;\hat{\mu},\hat{\sigma}^2+h^2)\right)^2 dx$$

Si se estandarizan los datos previamente ($y_i = (x_i - \hat{\mu})/\hat{\sigma}$) y se corrige la ventana del estimador de forma acorde (se toma $h/\hat{\sigma}$), el estadístico T puede expresarse como

$$T = \int_{\mathbb{R}} \left(\hat{f}_s(y) - f_N(x;0,\tau^2)\right)^2 dy,$$

donde $\hat{f}_s$ es el estimador núcleo construido a partir de los datos estandarizados y $\tau^2 = 1 + (h/\hat{\sigma})^2$. El valor de h adecuado en este caso es el dado por la regla de referencia a laa normal, h_N. Así,

$$\tau^2 = 1 + (h_N/\hat{\sigma})^2 = 1 + (1{,}059\hat{\sigma}n^{-1/5}/\hat{\sigma})^2 = 1 + (1{,}059n^{-1/5})^2.$$

La distribución de T bajo H_0 se puede estudiar teóricamente, pero resulta mucho más sencillo aproximarla mediante simulación.

En el siguiente ejemplo se utiliza la función `nise` de la librería `sm` (Bowman y Azzalini 1997) para llevar a cabo un contraste de normalidad (`nise` viene de *Normal Integrated Square Error*).

Ejemplo 3.12

Se realiza un contraste de normalidad para la variable `LSTAT` y `logitLSTAT <- log( (LSTAT/100) / 1 - (LSTAT/100))`.

```
   # Contraste de normalidad
nise.obs <- nise(LSTAT)*10000
logitLSTAT <- log( (LSTAT/100) / 1 - (LSTAT/100))
nise.obs.logit <- nise(logitLSTAT)*10000
n <- length(LSTAT)
S<-1000
sim.nise <- replicate(S, expr=nise(rnorm(n))*10000)
pval <- sum(sim.nise>nise.obs)/S
pval.logit <- sum(sim.nise>nise.obs.logit)/S
print(c(nise.obs,nise.obs.logit,quantile(sim.nise,.95),
        pval,pval.logit))
```

nise.obs	nise.obs.logit	quantile(sim.nise,.95)	pval	pval.logit
46.324905	4.281054	2.000750	0.000000	0.002000

3.6.3. Bandas de referencia normal

Una forma gráfica de contrastar normalidad es dibujar simultáneamente el estimador no paramétrico de la densidad $\hat{f}(x)$ y el estimador paramétrico máximo verosímil bajo normalidad, corregido por el sesgo de la estimación núcleo, $f_N(x; \hat{\mu}, \hat{\sigma}^2 + h^2)$.

El gráfico es aún más claro si se dibujan alrededor de $f_N(x; \hat{\mu}, \hat{\sigma}^2 + h^2)$ unas bandas que reflejen la variabilidad admisible en la estimación núcleo de la densidad si los datos realmente viniesen de una normal.

Ya vimos en la demostración del Teorema 3.3 (página 59) que si los datos son normales y se usa núcleo normal con desviación típica h, entonces

$$E(\hat{f}(x)) = f_N(x; \mu, \sigma^2 + h^2),$$

$$V(\hat{f}(x)) = \frac{1}{n}\left[f_N(0; 0, 2h^2) f_N(x; \mu, \sigma^2 + \frac{1}{2}h^2) - f_N(x; \mu, \sigma^2 + h^2)^2\right].$$

Así, las bandas de referencia normal serán

$$\left(\widehat{E(\hat{f}(x))} \mp z_{\alpha/2}\sqrt{\widehat{V(\hat{f}(x))}}\right).$$

En esta expresión se usan la media y la varianza muestrales como estimadores de μ y σ^2, respectivamente.

La función `sm.density` de la librería `sm` de `R` dibuja las bandas de referencia normal si se usa la opción `model="normal"`.

Ejemplo 3.13

La Figura 3.23 muestra las bandas de referencia a la normal para los estimadores de la densidad de las variables `LSTAT` (porcentaje de población con estatus social en la categoría inferior) y de esta variable transformada mediante la función logit:

```
logitLSTAT <- log( (LSTAT/100) / 1 - (LSTAT/100))
```

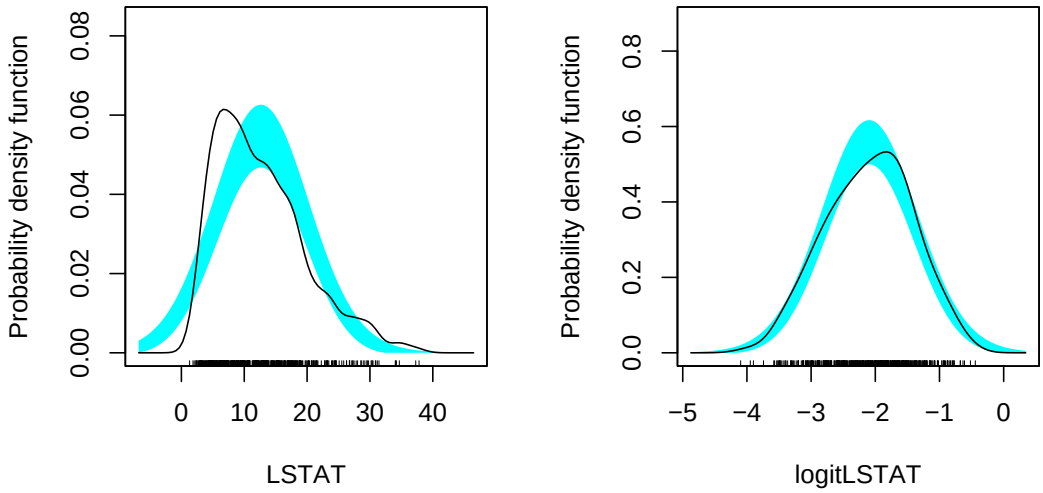


Figura 3.23: Estimadores núcleo de la densidad de las variables `LSTAT` y `log(LSTAT)`, acompañados de bandas de referencia a la normal.

3.6.4. Contraste de independencia

Sea $(x_1, y_1), \ldots, (x_n, y_n)$ una m.a.s. de $(X, Y) \sim f_{XY}(x, y)$. Se desea contrastar la independencia de X e Y:

$$\begin{cases} H_0 : X, Y \text{ independientes} \iff f_{XY}(x, y) = f_X(x) f_Y(y) \text{ para todo } x, y \\ H_1 : X, Y \text{ no son independientes} \iff \text{existen } x, y \text{ tales que } f_{XY}(x, y) \neq f_X(x) f_Y(y). \end{cases}$$

El estadístico de un contraste basado en la razón de verosimilitudes es

$$T = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\hat{f}_{XY}(x_i, y_i)}{\hat{f}_X(x_i)\hat{f}_Y(y_i)},$$

donde $\hat{f}_X$ y $\hat{f}_Y$ son los estimadores núcleo usuales y

$$\hat{f}_{XY}(x, y) = \frac{1}{nh_x h_y}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h_x}\right) K\left(\frac{y - y_i}{h_y}\right)$$

es un estimador núcleo bidimensional de la densidad conjunta $f_{XY}(x, y)$ (volveremos sobre este tipo de estimadores en la Sección 3.5).

Para calcular los valores críticos de T se puede usar un test de permutaciones:

1. Construir la muestra permutada

   $$(x_1, y_1^*), \ldots, (x_n, y_n^*)$$

   donde $(y_1^*, \ldots, y_n^*)$ es una permutación aleatoria de $(y_1, \ldots, y_n)$.

2. Calcular T^*, el valor del estadístico T en la muestra permutada. Bajo la hipótesis nula de independencia, la distribución de T y la de T^* coinciden.

3. Repetir los pasos 1 y 2 B veces: $T_1^*, \ldots, T_B^*$.

4. Calcular el p-valor del test como

   $$p - \text{valor} = \frac{\#\{T_b^* \geq T : b = 1 \ldots B\}}{B}.$$

3.6.5. Bootstrap en la estimación de la densidad

Sean $x_1, \ldots, x_n$ datos independientes idénticamente distribuidos generados a partir de $X \sim f(x)$. Sea $\hat{f}(x)$ el estimador núcleo de la densidad $f(x)$.

El procedimiento bootstrap que intenta imitar esa forma de generar datos aleatorios funciona como sigue:

1. Generar una muestra bootstrap mediante alguno de los dos siguientes esquemas:

   Bootstrap usual: $x_1^*, \ldots x_n^*$ i.i.d. según la distribución empírica de $x_1, \ldots, x_n$. Es decir, $x_1^*, \ldots x_n^*$ es una muestra de tamaño n tomada de la muestra original con reemplazamiento.

Bootstrap suavizado: $x_1^*, \dots x_n^*$ i.i.d. según $X^* \sim \hat{f}(x)$.

2. Calcular el estimador núcleo $\hat{f}^*(x)$ a partir de la muestra bootstrap.
3. Repetir los pasos anteriores B veces: $\hat{f}_1^*(x), \dots, \hat{f}_B^*(x)$.

Es de esperar que la distribución de $\hat{f}_b^*(x)$ alrededor de $\hat{f}(x)$ imite la distribución de $\hat{f}(x)$ alrededor de $f(x)$.

Si se usa el bootstrap usual se tiene que $E(\hat{f}_b^*(x)) = \hat{f}(x)$, es decir, no hay sesgo. Por tanto el bootstrap usual no imita bien el sesgo del estimador núcleo. En cambio, el bootstrap suavizado sí imita bien este sesgo.

Ambos procedimientos imitan bien la varianza del estimador núcleo. Por lo tanto, ambos procedimientos son útiles para construir bandas de variabilidad.

Usando bootstrap suavizado se pueden construir bandas de confianza puntuales. Sean

$$\overline{\hat{f}^*}(x) \text{ y } S^2(\hat{f}^*(x))$$

la media y la varianza muestrales de los B valores $\hat{f}_b^*(x)$. El sesgo del estimador $\hat{f}^*(x)$ se estima mediante

$$\overline{\hat{f}^*}(x) - \hat{f}(x)$$

y esta misma cantidad sirve como estimador de $\hat{f}(x)$.

Así, las bandas de confianza para $f(x)$ son

$$\left\{ \left(\hat{f}(x) - (\overline{\hat{f}^*}(x) - \hat{f}(x)) \right) \mp z_{\alpha/2} S(\hat{f}^*(x)) : x \in \mathbb{R} \right\}.$$

3.6.6. Contraste de igualdad de distribuciones

Sea $x_1, \dots, x_n$ una m.a.s. de $X \sim f_X(x)$, y sea $y_1, \dots, y_m$ una m.a.s. de $Y \sim f_Y(y)$. Se desea contrastar la igualdad de las distribuciones de X e Y:

$$\begin{cases} H_0 : f_X = f_Y \\ H_1 : f_X \neq f_Y \end{cases}$$

Un estadístico para este contraste puede ser

$$T = \int_{\mathbb{R}} (\hat{f}_X(u) - \hat{f}_Y(u))^2 du$$

donde $\hat{f}_X$ y $\hat{f}_Y$ son estimadores núcleo de las densidades de X e Y, respectivamente.

Para tabular la distribución de T bajo la hipótesis nula se pueden obtener pseudo-muestras generadas bajo el supuesto de igual distribución. Dos formas de hacer esto son las siguientes:

Bootstrap suavizado: Se estima una única densidad $\hat{f}$ a partir de las dos muestras conjuntamente. Se generan $n+m$ datos de esa densidad estimada. Los primeros n se asignan a una muestra y los m restantes a la otra.

Muestras permutadas: Se permutan aleatoriamente los $n+m$ datos. Los primeros n datos de la muestra permutada se asignan a una muestra y los m restantes a la otra.

Se puede hacer un contraste gráfico como sigue. Observar que

$$V(\sqrt{\hat{f}_X} - \sqrt{\hat{f}_Y}) = V(\sqrt{\hat{f}_X}) + V(\sqrt{\hat{f}_Y}) \approx \frac{1}{4}\frac{R(K)}{nh_n} + \frac{1}{4}\frac{R(K)}{nh_m}.$$

Si $n = m$,

$$V(\sqrt{\hat{f}_X} - \sqrt{\hat{f}_Y}) \approx \frac{1}{2}\frac{R(K)}{nh}.$$

Se construye el estimador promedio

$$\hat{f} = \frac{1}{2}(\hat{f}_X + \hat{f}_Y)$$

y se dibujan las bandas

$$\left\{ \left(\hat{f}(x) \mp z_{\alpha/2}\sqrt{\frac{R(K)\hat{f}(x)}{nh}} \right) : x \in \mathbb{R} \right\}.$$

Si los estimadores $\hat{f}_X$ y $\hat{f}_Y$ entran en las bandas podemos aceptar la hipótesis nula de igualdad de distribuciones.

La función `sm.density.compare` de la librería `sm` de `R` dibuja simultáneamente las bandas de aceptación de la hipótesis nula y los dos estimadores.

3.6.7. Discriminación no paramétrica basada en estimación de la densidad

En esta sección trataremos el PROBLEMA DE DISCRIMINACIÓN (o clasificación supervisada) desde una perspectiva no paramétrica. El planteamiento es el siguiente. Se observan p características, $x = (x_1, \ldots, x_p)$, en n individuos que pertenecen a una población dividida en q subpoblaciones (o clases), $\{C_1, \ldots, C_q\}$. De cada individuo también se sabe la clase $y_i \in \{1, \ldots, q\}$ a la que pertenece. Así, los datos de que se dispone son

$$(y_i; x_{i1}, \ldots, x_{ip}),\ i = 1, \ldots, n.$$

El objetivo es buscar una REGLA DISCRIMINANTE que asigne un nuevo individuo (cuya clase desconocemos) a una de las q clases, a partir de sus valores x_j.

La regla óptima, en el sentido de minimizar la probabilidad de mala clasificación, es la regla Bayes, que consiste en asignar el individuo con observaciones x a la clase j que tiene máxima probabilidad a posteriori:

$$f(x|C_j)P(C_j) = Max_{k=1\cdots q} f(x|C_k)P(C_k).$$

La regla Bayes sólo es aplicable si se conocen las probabilidades a priori de cada clase, $P(C_k)$, y las funciones de densidad $f(x|C_k)$ del conjunto de variables x para los individuos de cada clase. Las probabilidades a priori pueden estimarse fácilmente como las frecuencias relativas observadas de cada clase: $\hat{P}(C_k)$. La estimación de las funciones de densidad $f(x|C_k)$ puede llevarse a cabo mediante las técnicas ya estudiadas.

Al estimar no paramétricamente la densidad de x en cada una de las q clases en las que está dividida la población, se usan exclusivamente las observaciones que pertenecen a la clase cuya densidad se está estimando.

Finalmente, la regla discriminante consistirá en asignar el individuo con observaciones x a la clase j que tiene máxima probabilidad a posteriori estimada:

$$\arg \max_{k=1\cdots q} \hat{f}(x|C_k)\hat{P}(C_k).$$

Ejemplo 3.14

Apliquemos la técnica descrita al ejemplo que venimos usando en este capítulo: los datos sobre viviendas en los barrios de Boston. Dividimos los datos en dos clases, C_1 y C_2, según si la variable `RM` es menor o mayor que su mediana (que vale 6.2), respectivamente. Así las probabilidades a priori de cada clase serán iguales a 0.5. Como variables predictivas x tomaremos únicamente la variable `LSTAT` (modelo univariante). Se estima la densidad de `LSTAT` en cada clase usando transformación logarítmica, un núcleo gaussiano y ventanas elegidas por el método *plug-in*. La Figura 3.24 muestra las estimaciones obtenidas.

Se puede observar que la estimación de la densidad correspondiente a la clase C_1 es mayor que la correspondiente a C_2 si y sólo si `LSTAT` es mayor que 9.38. Por lo tanto la regla discriminante no paramétrica asignará a C_1 todas las observaciones con valores de `LSTAT` mayores que 9.38 y asignará las restantes a C_2.

.

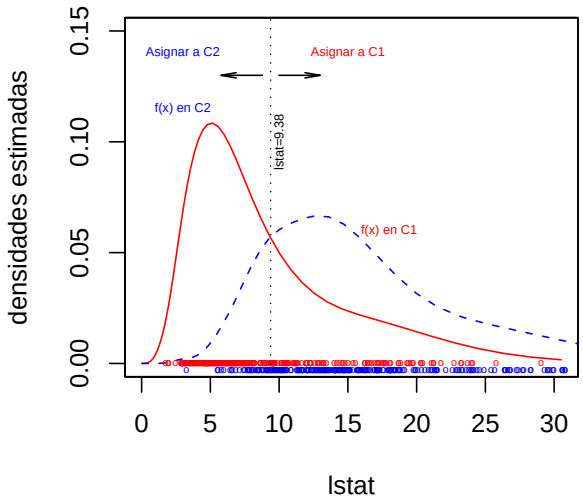


Figura 3.24: Estimaciones de las funciones de densidad de `LSTAT` en las clases C1 y C2 según la variable `RM`.

Ejemplo 3.15

En el ejemplo anterior, al sólo existir una sola variable explicativa, no se ve claramente la flexibilidad de la discriminación no paramétrica. Para ilustrar esta flexibilidad incluimos `AGE` como variable explicativa adicional. La variable `AGE` mide en cada barrio de Boston el porcentaje de viviendas construidas antes de 1940. Ahora son bivariantes las densidades que se han de estimar. La Figura 3.25 muestra la estimación de la densidad conjunta de (`LSTAT`,`AGE`) en cada una de las dos clases en las que se ha dividido la muestra. Se aprecia que la clase C_1 se concentra en valores relativamente altos de ambas variables (la moda está cerca del punto (`LSTAT`=15, `AGE`=90)), mientras que C_2 lo hace en valores bajos (la moda se encuentra en torno a (`LSTAT`=5, `AGE`=30)).

Para obtener la regla discriminante se toma la diferencia de la densidad estimada en la clase C_2 menos la estimada en C_1, y se clasifican en C_2 las observaciones para las que esta diferencia sea positiva. En la Figura 3.26 se ha representado esta diferencia y se ha señalado en trazo grueso la frontera entre las zonas que se clasificarán en una u otra clase, que es donde la diferencia entre las densidades estimadas es igual a 0. Se han marcado con un

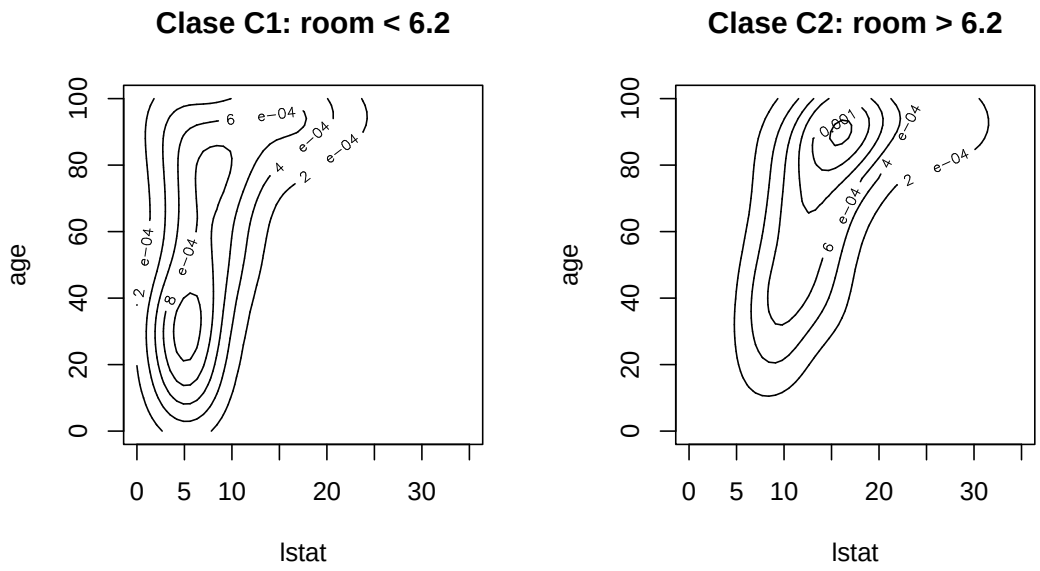


Figura 3.25: Estimación de la densidad de (`LSTAT`, `AGE`) en las clases C_1 y C_2.

círculo los puntos de C_2 y con una cruz los de C_1. Se ve claramente que los discriminadores obtenidos mediante estimación no paramétrica de la densidad pueden realizar clasificaciones no lineales.

.———————————————————————————————————————.

Recordemos que la estimación no paramétrica de la función de densidad sufre de la maldición de la dimensionalidad. Su efecto es menor sobre las reglas discriminantes derivadas de los estimadores de las densidades (porque para discriminar bien no es necesario estimar bien las funciones de densidad completas, sino sólo los valores relativos de las densidades en las distintas subpoblaciones) pero aún así, no es recomendable usar el método de discriminación descrito si la dimensionalidad p es grande (digamos mayor o igual que 4).

Una forma de solventar este problema es usar un estimador de la función de densidad construido bajo la hipótesis (poco verosímil en la mayor parte de los casos) de que las p componentes de las variables x son independientes. Así basta estimar no paramétricamente la densidad de cada variable explicativa y multiplicar éstas para obtener el estimador de la densidad conjunta.

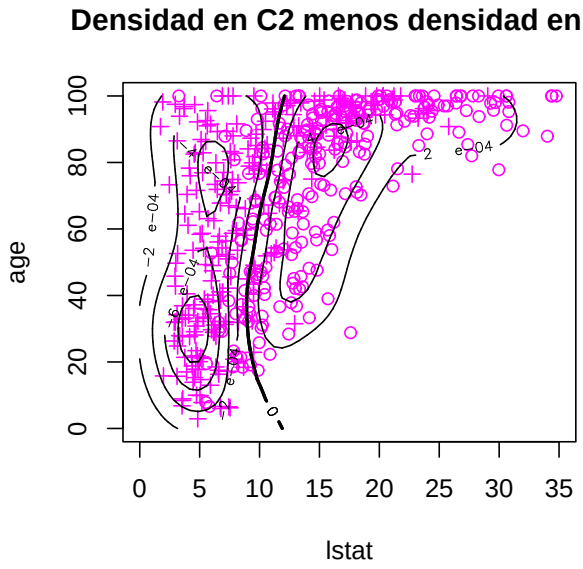


Figura 3.26: Diferencia de las densidades estimadas en las clases C_2 y C_1.

La regla discriminante obtenida a partir de ese estimador se conoce como REGLA BAYES NAIVE y en la práctica da buenos resultados. Para más detalles sobre este estimador puede consultarse la Sección 6.3.3. de Hastie, Tibshirani y Friedman (2001). Es evidente la analogía existente entre el paso de la estimación de la densidad multivariante al estimador basado en independencia de las marginales, y el paso del modelo de regresión múltiple no paramétrico al modelo aditivo, que estudiaremos en el Capítulo 6.

3.7. Otros estimadores de la densidad

3.7.1. Los k vecinos más cercanos

En la Sección 3.3.2, página 75, se presentó el ESTIMADOR NÚCLEO DE LOS k VECINOS MÁS CERCANOS como caso particular de los estimadores núcleo locales.

Es posible motivar este estimador directamente a partir de la relación entre la función de densidad $f(x)$ y la función de distribución $F(x)$, que estimaremos mediante la función de distribución empírica $\hat{F}$:

$$f(x) = \frac{d}{dx}F(x) = \lim_{h \longrightarrow 0} \frac{F(x+h) - F(x-h)}{2h}$$

$$\approx \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h} = \frac{k(x,h)/n}{2h},$$

donde se ha usado la notación $k(x,h) = \#\{x_i : x_i \in [x-h, x+h]\}$. Fijar un valor de h equivale a fijar el denominador de la última fracción y dejar que el numerador dependa de x.

También podríamos estimar la derivada de F en x fijando el numerador y dejando que el denominador dependa de x:

$$\hat{f}_k(x) = \frac{k/n}{2h(x,k)},$$

con $h(x,k)$ igual a la distancia a x de su k-ésimo vecino más cercano.

Es fácil probar que

$$\hat{f}_k(x) = \frac{1}{nh(x,k)} \sum_{i=1}^{n} K_U\left(\frac{1}{nh(x,k)}\right),$$

donde K_U es el núcleo uniforme en $[-1,1]$. Si se sustituye K_U por otro núcleo K se obtiene un estimador que hereda las propiedades de suavidad de K. Así, la expresión genérica de un estimador de la densidad por k vecinos más cercanos es

$$\hat{f}_k(x) = \frac{1}{nh(x,k)} \sum_{i=1}^{n} K\left(\frac{1}{nh(x,k)}\right).$$

Observar que $f(x) \approx \hat{f}_k(x) = k/(2nh(x,k))$ implica que $h(x,k) \approx k/(2nf(x))$. Es decir, el estimador de los k vecinos más cercanos es un estimador núcleo local con $h(x) = h(x,k) \propto 1/f(x)$.

El estimador $\hat{f}_k(x)$ no necesariamente es una función de densidad. Puede ser muy abrupto en algunos valores de x.

El parámetro de suavizado de este estimador es k, el número de vecinos involucrados en la estimación de $f(x)$. Para que $\hat{f}_k(x)$ converja a $f(x)$ es condición suficiente k dependa de n de forma que

$$\lim_n k(n) = \infty, \ \lim_n \frac{k(n)}{n} = 0.$$

Se puede probar que para un x fijo el valor de k óptimo (en el sentido del MSE) es $k(n) = O(n^{4/5})$.

La selección automática del parámetro de suavizado k se complica en este caso porque el IMSE es en general no acotado. Se puede recurrir a la validación cruzada basada en máxima verosimilitud.

3.7.2. Desarrollos en series de funciones ortogonales

Supongamos que X tiene soporte en $[-\pi, \pi]$ y que su función de densidad es de cuadrado integrable en ese intervalo:

$$f \in \mathcal{L}_2([-\pi,\pi]) = \{\phi : [-\pi,\pi] \longrightarrow \mathbb{R} \text{ tal que } \frac{1}{2\pi}\int_{-\pi}^{\pi} \phi(x)^2 dx < \infty\}.$$

Ese espacio de funciones $\mathcal{L}_2([-\pi,\pi])$ es un espacio vectorial euclídeo (tiene producto escalar: $< f, g >= (1/(2\pi))\int_{-\pi}^{\pi} f(x)g(x)dx$) y una de sus bases ortonormales la constituyen las funciones trigonométricas

$$\{1, \cos(x), \sin(x), \cos(2x), \sin(2x), \ldots\}.$$

Por lo tanto, esa función de densidad f tiene un desarrollo en esa base de funciones ortonormales, que se conoce como DESARROLLO EN SERIE DE FOURIER de f:

$$f(x) = a_0 + \sum_{k=1}^{\infty}(a_k \cos(kx) + b_k \sin(kx)).$$

Cada coeficiente del desarrollo se puede calcular como

$$a_0 =< f(x), 1 >= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)dx = \frac{1}{2\pi}.$$

$$a_k =< f(x), \cos(kx) >= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\cos(kx)dx = \frac{1}{2\pi}E(\cos(kX)),$$

$$b_k =< f(x), \sin(kx) >= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\sin(kx)dx = \frac{1}{2\pi}E(\sin(kX)).$$

Los coeficientes pueden estimarse a partir de la muestra de X observada:

$$\hat{a}_k = \frac{1}{2\pi}\frac{1}{n}\sum_{i=1}^{n}\cos(kx_i),\ \hat{b}_k = \frac{1}{2\pi}\frac{1}{n}\sum_{i=1}^{n}\sin(kx_i).$$

Si se trunca el desarrollo de f en el término m y se sustituyen los coeficientes desconocidos por sus estimaciones, se tiene un estimador de la función de densidad:

$$\hat{f}_m(x) = \frac{1}{2\pi} + \sum_{k=1}^{m}(\hat{a}_k \cos(kx)/\pi + \hat{b}_k \sin(kx)/\pi).$$

El parámetro de suavizado aquí es m. Para tener convergencia puntual del estimador al valor de la función de densidad es suficiente que $m = m(n)$ vaya a infinito y que m/n vaya a 0 cuando n tiende a infinito.

Existen otras bases de funciones ortonormales no restringidas a $[-\pi, \pi]$. Volveremos sobre ello en el Capítulo 5.

3.7.3. Máxima verosimilitud penalizada

El problema de la estimación de la densidad se podría plantear como un problema de estimación paramétrica, con la particularidad de que el espacio paramétrico tiene dimensión infinita. En efecto, sea $\mathbb{R}^{\mathbb{R}}$ el conjunto de funciones de $\mathbb{R}$ en $\mathbb{R}$. Se define el espacio paramétrico

$$\Theta = \{f \in \mathbb{R}^{\mathbb{R}} : f(x) \geq 0 \ \forall x \in \mathbb{R}, \int_{\mathbb{R}} f(x)dx = 1\}.$$

Consideramos la v.a. X con densidad $f \in \Theta$ y una realización de una m.a.s. de X: $x_1, \ldots, x_n$. Estimar la densidad de X equivale a estimar el parámetro f. Observar que la dimensión de Θ no es finita.

En este contexto, el estimador máximo verosímil de θ no está bien definido porque la función de verosimilitud no está acotada en $\theta \in \Theta$. En efecto, la función de verosimilitud y su logaritmo evaluadas en $g \in \Theta$ son

$$L(g; x_1, \ldots, x_n) = \prod_{i=1}^{n} g(x_i), \ l(g; x_1, \ldots, x_n) = \sum_{i=1}^{n} \log g(x_i).$$

La función de verosimilitud es no acotada,

$$\sup_{g \in \Theta} L(g; x_1, \ldots, x_n) = \infty,$$

porque si tomamos como g

$$g_\varepsilon(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2\varepsilon} I_{[-\varepsilon, \varepsilon]}(x)$$

tenemos que si ε es menor que la menor de las diferencias entre dos datos x_i consecutivos,

$$L(g_\varepsilon; x_1, \ldots, x_n) = \frac{1}{n^n 2^n \varepsilon^n}$$

y

$$\lim_{\varepsilon \longrightarrow 0} L(g_\varepsilon; x_1, \ldots, x_n) = \infty.$$

Las funciones $g \in \Theta$ que hacen que $L(g_\varepsilon; x_1, \ldots, x_n)$ tome valores grandes son muy abruptas, porque tienen picos (máximos locales) muy pronunciados en algunos o en todos los valores x_i observados.

Una alternativa a la estimación máximo verosímil es realizar una maximización de la verosimilitud penalizando la falta de suavidad de las funciones

candidatas a máximo. Se tiene así el problema de la máxima verosimilitud penalizada:

$$\max_{g\in\Theta_\Phi} \sum_{i=1}^{n} \log g(x_i) - \lambda\Phi(\log g),$$

donde $\Phi(\phi)$ es un término que mide la variabilidad de ϕ y $\Theta_\Phi = \{g \in \Theta : \Phi(\log g)$ está bien definido$\}$.

Por ejemplo, $\Phi(\phi) = \int_{\mathbb{R}} (\phi''(x))^2 dx$ es una de las medidas de variabilidad que pueden ser usadas.

Este procedimiento expresa explícitamente lo que buscamos al estimar la densidad no paramétricamente: maximizar la verosimilitud descartando funciones demasiado abruptas.

Si usamos la notación

$$\phi(x) = \log g(x) \Longleftrightarrow g(x) = \frac{e^{\phi(x)}}{\int_{\mathbb{R}} e^{\phi(u)} du}$$

el problema anterior es equivalente a este otro:

$$\max_{\phi:\mathbb{R}\longrightarrow\mathbb{R}} \left\{ \sum_{i=1}^{n} \phi(x_i) - \lambda \int_{\mathbb{R}} (\phi''(x))^2 - n \log \int_{\mathbb{R}} e^{\phi(u)} du \right\}.$$

Si $\hat{\phi}$ es el óptimo, entonces el estimador no paramétrico de la densidad de X es

$$\frac{e^{\hat{\phi}(x)}}{\int_{\mathbb{R}} e^{\hat{\phi}(u)} du}$$

Se puede probar que la función óptima $\hat{\phi}$ es una función spline de tercer grado (un spline cúbico) con nodos en los puntos observados $x_{(1)}, \ldots, x_{(n)}$.

Un spline cúbico es una función que a trozos (entre $x_{(i)}$ y $x_{(i+1)}$) es un polinomio de tercer grado, y en los puntos $x_{(i)}$, las expresiones polinómicas que definen la función a un lado y a otro enlazan bien, en el sentido de que en esos puntos la función es continua y tiene sus dos primeras derivadas continuas.

El parámetro de suavizado es aquí el parámetro λ que penaliza la falta de suavidad. Si λ es grande la función estimada es suave, mientras que si λ es pequeño el estimador es más abrupto. Si λ tiende a infinito, se puede probar que $\hat{\phi}$ tiende a un polinomio de primer grado y, por tanto, si se supone soporte $[0, \infty)$ la estimación de la densidad tiende a ser la densidad de una exponencial con parámetro λ igual a su estimador máximo verosímil.

Volveremos a tratar la estimación de funciones mediante splines en el Capítulo 5.

3.7.4. Verosimilitud local

Se trata de una técnica que consiste en proponer un modelo paramétrico alrededor de cada punto x donde se desea estimar la densidad. Este modelo se estima por máxima verosimilitud (localmente, alrededor de x). Después se enlazan las estimaciones paramétricas locales para obtener un estimador no paramétrico global de la densidad. Puede verse el artículo Delicado (2006) y las referencias incluidas en él.

En la sección 4.4 volveremos a tratar el problema de estimar modelos no paramétricos mediante la formulación de modelos paramétricos localmente válidos. Será entonces momento de profundizar en la estimación de la densidad por verosimilitud local.

3.7.5. Representación general

Bajo condiciones no muy restrictivas se puede probar que cualquier estimador no paramétrico de la densidad puede escribirse como un estimador núcleo generalizado:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(x,x_i)} K_x\left(\frac{x - x_i}{h(x,x_i)}\right).$$

3.8. Software

3.8.1. Estimación de la densidad en R

El paquete estadístico de distribución libre R (`http://www.r-project.org/`, `http://cran.r-project.org/`). Aquí puede usarse la función incluida por defecto para estimación de la densidad (`density`) o la librería `sm` que acompaña el libro Bowman y Azzalini (1997) (en concreto, la función `sm.density` y otras relacionadas con ella).

3.8.2. Estimación de la densidad en MATLAB

El Profesor Christian Beardah (School of Biomedical and Natural Sciences, Nottingham Trent University, UK) escribió una *toolbox* en MATLAB llamada KDE (Kernel Density Estimation) en los años 90 del pasado siglo. Más tarde estas funciones sirvieron de base a la toolbox kdetools desarrollada en la sección de Mathematical Statistics del Centre for Mathematical Sciences de la Universidad de Lund (Suecia). Esta librería es parte de una toolbox más amplia llamada WAFO, que puede descargarse en `http://www.maths.lth.se/matstat/wafo/`.

Capítulo 4

Estimación de la función de regresión

REFERENCIAS: Wand y Jones (1995), Simonoff (1996), Fan y Gijbels (1996), Bowman y Azzalini (1997), Wasserman (2006).

En este capítulo estudiaremos la predicción no paramétrica del valor esperado de una variable aleatoria *dependiente* (Y) condicionando a que otras variables *predictoras* (X) toman unos valores conocidos (x), lo que se conoce como FUNCIÓN DE REGRESIÓN:

$$m(x) = E(Y|X = x).$$

Habitualmente estas predicciones se derivarán de la propuesta y posterior estimación de modelos para la distribución condicionada $(Y|X = x)$.

Los MODELOS DE REGRESIÓN PARAMÉTRICOS suponen que los datos observados provienen de variables aleatorias cuya distribución es conocida, salvo por la presencia de algunos parámetros cuyo valor se desconoce. Por ejemplo, la relación entre el peso (y) y la altura (x) de un grupo de personas puede modelarse mediante regresión lineal con errores normales:

$$y = \beta_0 + \beta_1 x + \varepsilon, \text{ con } \varepsilon \sim N(0, \sigma^2).$$

Éste es un modelo estadístico con tres parámetros desconocidos: β_0, β_1 y σ^2.

Una formulación general de un modelo de regresión paramétrico es la siguiente:

$$y_i = m(x_i; \theta) + \varepsilon_i,\ i = 1, \ldots, n,\ \theta \in \Theta \subseteq \mathbb{R}^p,$$

donde $m(x; \theta)$ es una función conocida de x y θ, que es desconocido, $\varepsilon_1, \ldots, \varepsilon_n$ son v.a.i.i.d. con $E(\varepsilon_i) = 0$ y $V(\varepsilon_i) = \sigma^2$. El modelo de regresión lineal simple es un caso particular con $\theta = (\beta_0, \beta_1)$ y $m(x; \beta_0, \beta_1) = \beta_0 + \beta_1 x$.

Se dice que se ajusta el modelo paramétrico cuando se estiman sus parámetros a partir de un conjunto de observaciones que siguen dicho modelo. Si el modelo $m(x;\theta)$ es correcto, la estimación de los parámetros (θ, σ^2) puede realizarse con una cantidad pequeña de datos (por ejemplo, puede usarse el método de mínimos cuadrados o el de máxima verosimilitud), pueden hacerse predicciones de nuevos valores de y conocido el valor de x, y tener información precisa acerca de la incertidumbre asociada a la estimación y a la predicción. Éstas son algunas de las buenas propiedades de los modelos paramétricos. Además, en muchas ocasiones los parámetros tienen una interpretación intuitiva en términos relacionados con el problema en estudio (por ejemplo, β_1 es la derivada de y respecto de x en el modelo de regresión anterior).

Sin embargo, si el modelo paramétrico no es adecuado puede ser peor tenerlo ajustado que no tener nada, porque el modelo paramétrico conlleva un grado de exactitud en las afirmaciones que de él se derivan que son adecuadas cuando el modelo es correcto, pero que en caso contrario pueden estar muy alejadas de la realidad.

Los modelos paramétricos presentan un problema fundamental: su estructura es tan rígida que no pueden adaptarse a muchos conjuntos de datos.

Ejemplo 4.1

En el Ejemplo 2.2 se consideraba el conjunto de datos referido a la vivienda en 506 barrios de Boston en 1978 (Boston Housing Data). Ahí se observa que la relación entre las variables `RM` (número medio de habitaciones por vivienda) y `LSTAT` (porcentaje de población con estatus social en la categoría inferior) no se puede modelizar ni con un modelo de regresión lineal ni con uno cuadrático (ver Figura 4.1).

.

En este capítulo presentaremos una alternativa no paramétrica a los modelos de regresión paramétricos usuales. Por ejemplo, la relación entre el peso y la altura de una persona podría modelizarse no paramétricamente diciendo que

$$y = m(x) + \varepsilon,$$

donde $m(x)$ es una función (posiblemente continua o derivable) cuya forma no se especifica, y ε es una variable aleatoria con valor esperado igual a cero.

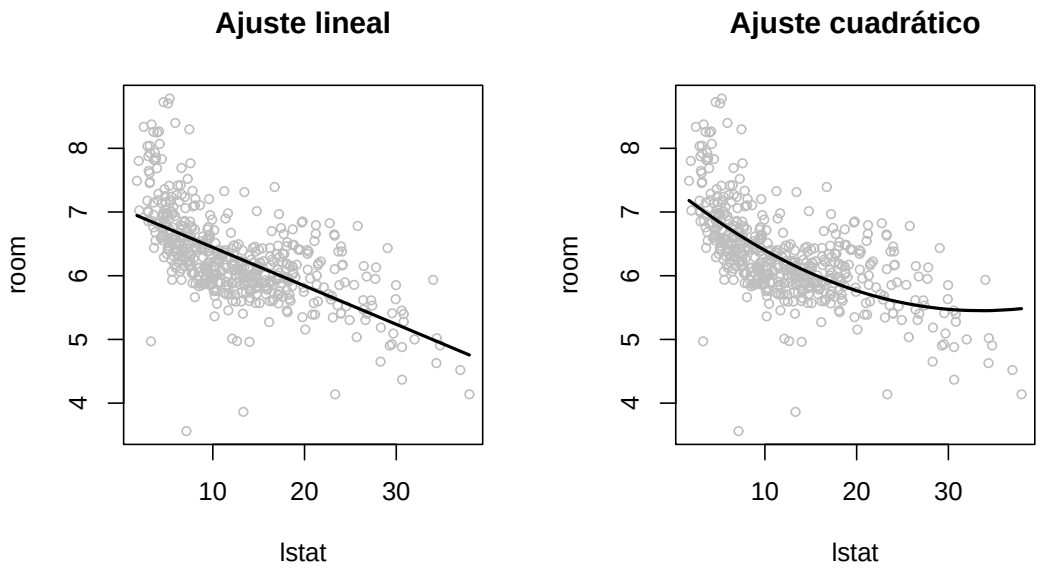


Figura 4.1: Ajustes paramétricos de la variable `RM` como función de la variable `LSTAT`.

4.1. El modelo de regresión no paramétrica

Comencemos en el contexto del modelo de regresión simple: la variable respuesta y es continua y sólo hay una variable explicativa x, también continua (el caso de la regresión múltiple lo abordaremos en el Capítulo 6). Se supone que se observan n pares de datos (x_i, y_i) que provienen del siguiente modelo de regresión no paramétrico:

$$y_i = m(x_i) + \varepsilon_i,\ i = 1, \ldots, n, \tag{4.1}$$

donde $\varepsilon_1, \ldots, \varepsilon_n$ son v.a. independientes con

$$E(\varepsilon_i) = 0, V(\varepsilon_i) = \sigma^2 \text{ para todo } i,$$

y los valores de la variable explicativa $x_1, \ldots, x_n$ son conocidos, por lo que se dice que el modelo tiene DISEÑO FIJO.

Dado que la varianza de los errores ε_i es constante diremos que el modelo es HOMOCEDÁSTICO. Esta hipótesis puede relajarse y suponerse que esa varianza es función de la variable explicativa x: $V(\varepsilon_i) = \sigma^2(x_i)$. En ese caso diremos que el modelo es HETEROCEDÁSTICO.

No se especifica la forma funcional de la función de regresión $m(x)$, aunque sí se supone que es una función suficientemente regular (por ejemplo, es habitual la hipótesis de que $m(x)$ tiene segunda derivada continua).

También puede plantearse el modelo de regresión con DISEÑO ALEATORIO. Sea (X, Y) v.a. bivariante con densidad conjunta $f(x, y)$. Se define la FUNCIÓN DE REGRESIÓN como $m(x) = E(Y|X = x)$. Entonces $E(Y|X) = m(X)$. Así, si definimos $\varepsilon = Y - m(X)$, se tiene que

$$Y = m(X) + \varepsilon,\ E(\varepsilon|X) = 0,\ V(\varepsilon|X) = \sigma^2(X).$$

Sean (X_i, Y_i), $i = 1, \ldots, n$ una m.a.s. de (X, Y). Estos datos siguen el modelo de regresión no paramétrico

$$Y_i = m(X_i) + \varepsilon_i,\ i = 1, \ldots, n.$$

Dado que en el modelo de regresión el interés es estudiar la esperanza de Y condicionada a que X toma un valor conocido x, supondremos que tenemos siempre un diseño fijo.

Si necesitamos en algún desarrollo una función de densidad $f(x)$ asociada a la variable explicativa (sería la densidad de X en diseño aleatorio) supondremos que los valores $x_1, \ldots, x_n$ provienen de un DISEÑO REGULAR de densidad $f(x)$. Es decir, los x_i se han obtenido así:

$$x_i = F^{-1}\left(\frac{i - 1/2}{n}\right), i = 1, \ldots, n, \ \text{ con } F(x) = \int_{-\infty}^{x} f(u)du.$$

Una vez establecido el modelo, el paso siguiente consiste en estimarlo (o ajustarlo) a partir de las n observaciones disponibles. Es decir, hay que construir un estimador $\hat{m}(x)$ de la función de regresión y un estimador $\hat{\sigma}^2$ de la varianza del error. Los procedimientos de estimación de $m(x)$ también se conocen como MÉTODOS DE SUAVIZADO (*smoothing* en inglés).

El abanico de técnicas disponibles para estimar no paramétricamente la función de regresión es amplísimo e incluye, entre otras, las siguientes:

- *Ajuste local de modelos paramétricos.* Se basa en hacer varios (o incluso infinitos, desde un punto de vista teórico) ajustes paramétricos teniendo en cuenta únicamente los datos cercanos al punto donde se desea estimar la función. Son las que desarrollaremos en este capítulo.

- *Métodos basados en series ortogonales de funciones.* Se elige una base ortonormal del espacio vectorial de funciones y se estiman los coeficientes del desarrollo en esa base de la función de regresión. Los ajustes por

series de Fourier y mediante *wavelets* son los dos enfoques más utilizados. En el Capítulo 5 trataremos este tema más ampliamente. También puede consultarse la Sección 2.5 de Fan y Gijbels (1996), el Capítulo 8 de Wasserman (2006) y las referencias allí citadas.

- *Suavizado mediante splines.* Se plantea el problema de buscar la función $\hat{m}(x)$ que minimiza la suma de los cuadrados de los errores ($e_i = y_i - \hat{m}(x_i)$) más un término que penaliza la falta de suavidad de las funciones $\hat{m}(x)$ candidatas (en términos de la integral del cuadrado de su derivada segunda). Se puede probar que la solución es un *spline* cúbico con nodos en los puntos x_i observados. En el Capítulo 5 se trata a fondo este método de estimación. Véase también el libro de Green y Silverman (1994).

- *Técnicas de aprendizaje supervisado.* Las redes neuronales, los k vecinos más cercanos y los árboles de regresión se usan habitualmente para estimar $m(x)$. De hecho cualquier técnica de aprendizaje supervisado que admita respuesta continua y predictor continuo puede usarse para estimar no paramétricamente la función de regresión. Para una visión de estos métodos desde un punto de vista estadístico pueden consultarse los capítulos 9, 11 y 13 de Hastie, Tibshirani y Friedman (2001).

En gran parte, este abanico de técnicas de regresión también puede ampliarse al problema de discriminación (en el que la respuesta es categórica), tal como se verá en la Sección 4.4.

4.2. Estimadores núcleo y polinomios locales

Retomamos el ejemplo de los datos de los barrios de Boston. La observación de la nube de puntos formada por las observaciones de las variables `LSTAT` y `RM`, así como del resultado del ajuste de la regresión lineal simple a estos datos mostrada en la Figura 4.1, sugieren que un único modelo lineal no es válido para todo el rango de la variable explicativa `LSTAT`. La primera idea que surge para solventar ese problema es dividir el rango de esta variable en intervalos, de forma que la relación entre las dos variables sea aproximadamente lineal en cada intervalo. Así, parece razonable considerar los cuatro intervalos delimitados por los valores 10 %, 20 % y 30 % de la variable `LSTAT`. Hecho esto, en cada intervalo se ajusta un modelo de regresión lineal simple y se obtiene el resultado que muestra el gráfico de la derecha en la Figura 4.2. El cálculo de la media muestral en cada uno de esos tramos daría lugar

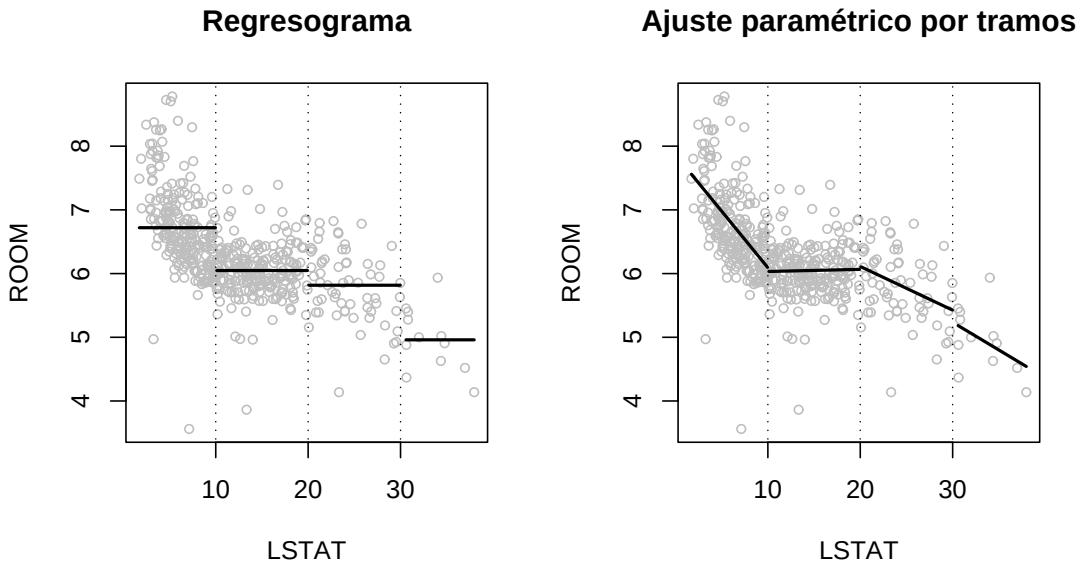


Figura 4.2: Regresograma de `RM` como función de `LSTAT` (izquierda) y ajuste de modelos de regresión lineal simple en cuatro intervalos del recorrido de la variable `LSTAT` (derecha).

a lo que se conoce como REGRESOGRAMA, por su analogía con el histograma (ver gráfico de la izquierda en la Figura 4.2).

Este sencillo ejercicio refleja todavía más claramente que los datos sugieren una relación entre las dos variables que no se ajusta a una regresión simple. Sin embargo el resultado obtenido dista mucho de ser completamente satisfactorio por varios motivos. En primer lugar, la función de regresión estimada mediante este procedimiento es discontinua en los puntos que definen los intervalos. En segundo lugar, en la estimación del valor de la función de regresión en puntos cercanos a los extremos de los intervalos (por ejemplo, en `LSTAT`=9) no intervienen muchos datos cercanos a ese valor de la variable explicativa y que pertenecen a un intervalo distinto (no se usan los datos para los que `LSTAT` está entre 10 y 14, por ejemplo) mientras que esa estimación sí se ve afectada por algunos datos del mismo intervalo que están más lejos (en nuestro ejemplo los datos para los que `LSTAT` es menor que 4).

Una forma de atacar la segunda de las deficiencias mencionadas es la siguiente. Para estimar la función de regresión en un valor concreto t de la variable explicativa, se debería usar un intervalo de la variable explicativa específico para ese valor t, centrado en el valor t y que sólo contenga datos en los que la variable explicativa tome valores cercanos a t. Así, si se desea estimar la función de regresión en el valor `LSTAT`=9, se usarán únicamente las observaciones para las cuales $4 <$ `LSTAT` < 14 (si es que queremos que los nuevos intervalos sigan teniendo 10 unidades de amplitud). Este procedimiento se ilustra en el panel izquierdo de la Figura 4.3.

Pese a la apariencia de continuidad de la función de regresión representada en este gráfico, el método descrito no proporciona estimadores continuos. Ello se debe a que, al desplazar a la derecha el intervalo que determina los datos activos para el cálculo de la regresión simple local, habrá puntos que dejen de ser activos (los situados más a la izquierda) y otros que pasen a serlo (los que estaban cerca del intervalo en su parte derecha). El hecho de que un dato pase de ser activo a no serlo, y viceversa, de forma abrupta (su peso en la regresión simple pasa de ser 0 a ser 1) hace que la función de regresión estimada no sea continua.

La continuidad del estimador puede conseguirse si se ponderan los datos de forma que el peso de una observación (x_i, y_i) sea función decreciente (que tienda a 0 y sea continua) de la distancia de su ordenada x_i al punto t donde se realiza la estimación. De esta forma, al desplazar el punto t, las observaciones irán tomando todos los valores posibles de la función peso de forma continua en t y, como resultado, se tendrá un estimador continuo de la función de regresión. Esto se ilustra en el panel derecho de la Figura 4.

La forma usual de asignar estos pesos es mediante una función núcleo (*kernel*) K (función simétrica no negativa, continua, decreciente en $[0, \infty)$ y

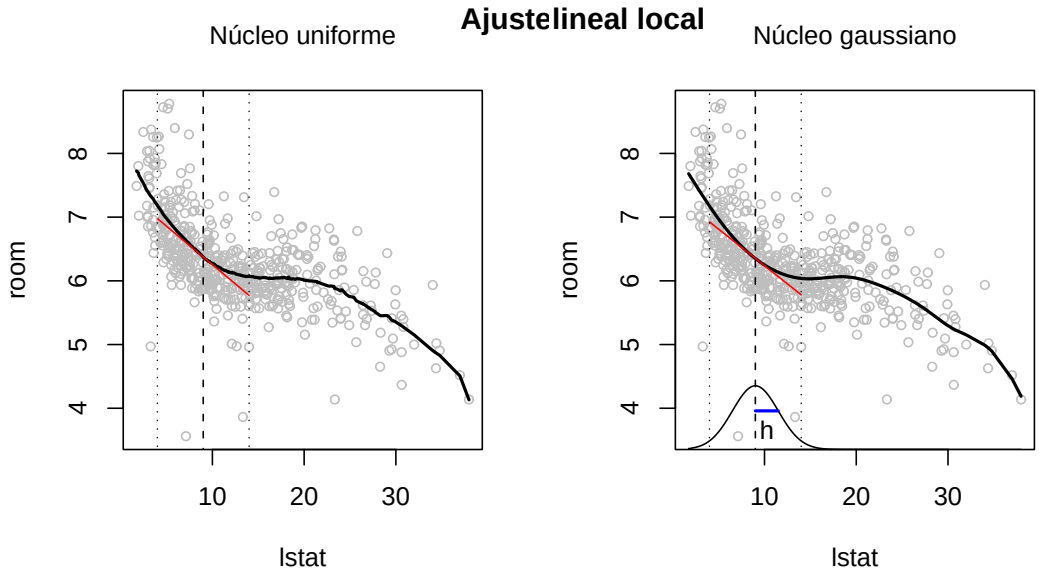


Figura 4.3: Ajuste lineal local en el punto `LSTAT`=9. Ponderación uniforme (izda.). Con núcleo gaussiano (dcha.).

que tiende a 0 cuando el argumento tiende a infinito). El peso de (x_i, y_i) en la estimación de $m(t)$ será

$$w_i = w(t, x_i) = \frac{K\left(\frac{x_i - t}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x_j - t}{h}\right)},$$

donde h es un parámetro de escala que controla la concentración del peso total alrededor de t: si h es pequeño únicamente las observaciones más cercanas a t tendrán peso relevante, mientras que valores grandes de h permiten que observaciones más alejadas de t también intervengan en la estimación de $m(t)$. A h se le denomina PARÁMETRO DE SUAVIZADO (o VENTANA) del estimador no paramétrico y permite controlar el grado de localidad (o globalidad) de la estimación. La estimación final se ve notablemente afectada por cambios en la elección del parámetro de suavizado, por lo que esta tarea resulta crucial en la estimación no paramétrica. A la elección de h dedicaremos la Sección 4.3.

Una vez determinados los pesos w_i, se resuelve el problema de mínimos cuadrados ponderados siguiente:

$$\min_{a,b} \sum_{i=1}^{n} w_i \left(y_i - (a + b(x_i - t))\right)^2.$$

Los parámetros a y b así obtenidos dependen de t, porque los pesos w_i dependen de t: $a = a(t), b = b(t)$. La recta de regresión ajustada localmente alrededor de t es

$$l_t(x) = a(t) + b(t)(x - t),$$

y la estimación de la función de regresión en el punto t es el valor que toma esa recta en x=t:

$$\hat{m}(t) = l_t(t) = a(t).$$

Ejemplo 4.2

El panel derecho de la Figura 4.3 muestra el estimador así construido para los datos (`LSTAT`, `RM`). Se indica también la recta de regresión estimada por mínimos cuadrados ponderados en el punto `LSTAT`=9. En este caso se ha usado como función núcleo K la función densidad de la variable aleatoria normal estándar, que se conoce como núcleo gaussiano. Se ha representado el núcleo en el mismo gráfico para ilustrar cómo se asignan los pesos a los distintos datos según la proximidad a 9 de los valores de la variable `LSTAT`. Se ha señalado también el valor $h = 2{,}5$ del parámetro de suavizado.

Las funciones núcleo usadas en estimación no paramétrica de la regresión son las mismas que las utilizadas en la estimación de la densidad (ver las Tablas 3.1 y 3.3 y las Figuras 3.10 y 3.11).

Obsérvese que usar un núcleo uniforme es equivalente a estimar la regresión localmente usando únicamente los puntos que están en un intervalo centrado en el punto donde se realiza la estimación, todos ellos con idéntico peso. Es decir, el procedimiento que seguimos para construir el estimador de la función de regresión representada en el panel izquierdo de la Figura 4.3 es el mismo que si usamos un núcleo uniforme con ventana $h = 5$.

El estimador lineal local se generaliza fácilmente al ajuste local de regresiones *polinómicas* de mayor grado. Una regresión polinómica es una regresión lineal múltiple en la que se incluyen variables que son potencias de la variable explicativa. Es decir, de una variable x obtenemos el polinomio

$$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_q x^q$$

y se actúa como en una regresión lineal múltiple con q regresores. En nuestro caso, en vez del valor de x_i, utilizaremos el valor $(x_i - t)$. A partir de aquí, el estimador de polinomios locales de grado q se construye como sigue. Primero se asignan los pesos w_i mediante una función núcleo, tal como se hace en el ajuste lineal local. Se plantea entonces el problema de regresión polinómica ponderada

$$\min_{\beta_0,\ldots,\beta_q} \sum_{i=1}^{n} w_i \left(y_i - (\beta_0 + \beta_1(x_i - t) + \cdots + \beta_q(x_i - t)^q)\right)^2 .$$

Obsérvese que los coeficientes obtenidos dependen del punto t donde se realiza la estimación: $\hat{\beta}_j = \hat{\beta}_j(t)$. Finalmente, se da como estimador de $m(t)$ el valor del polinomio $P_{q,t}(x - t) = \sum_{j=0}^{p} \hat{\beta}_j (x - t)^j$ estimado localmente en torno a $x = t$:

$$\hat{m}_q(t) = P_{q,t}(0) = \hat{\beta}_0(t).$$

El hecho de ajustar polinomios de grado mayor que 1 permite que la función estimada se ajuste mejor a los datos.

Además, a partir del polinomio $P_{q,t}(x - t)$ estimado alrededor de t se pueden dar estimadores de las primeras q derivadas de la función m en t. Por ejemplo, la derivada s-ésima de m en $x = t$ se estima como

$$\hat{m}_q^{(s)}(t) = \frac{d^s}{dx^s} \left(P_{q,t}(x - t)\right)\Big|_{x=t} = s!\hat{\beta}_s(t). \tag{4.2}$$

En el caso particular de que se ajuste localmente un polinomio de grado 0 (es decir una constante), se obtiene el conocido como *estimador de Nadaraya-Watson* o *estimador núcleo de la regresión*. Su expresión explícita es ésta:

$$\hat{m}_K(t) = \frac{\sum_{i=1}^n K\left(\frac{x_i - t}{h}\right) y_i}{\sum_{i=1}^n K\left(\frac{x_i - t}{h}\right)} = \sum_{i=1}^n w(t, x_i) y_i.$$

Históricamente el estimador de Nadaraya-Watson es anterior a los estimadores por polinomios locales. Obsérvese que $\hat{m}_K(t)$ es una media ponderada de los valores de la variable respuesta, donde el peso de cada dato depende de la distancia entre el valor de la variable explicativa y el punto t donde se está estimando la función de regresión. Podemos ver el estimador núcleo como una media ponderada móvil. De hecho, puede probarse que todo estimador por polinomios locales puede expresarse como una media ponderada,

$$\hat{m}_q(t) = \sum_{i=1}^n w_q^*(t, x_i) y_i.$$

aunque los pesos $w_q^*(t, x_i)$ no necesariamente han de ser positivos.

4.2.1. Derivación directa del estimador núcleo de la regresión

El modelo de regresión no paramétrica con diseño aleatorio, en el que la función de regresión es

$$m(x) = E(Y|X = x) = \int_{\mathbb{R}} y f_Y(y|X = x) dy = \int_{\mathbb{R}} y \frac{f(x, y)}{f_X(x)} dy,$$

sugiere un estimador de $m(x)$ obtenido a partir de los estimadores no paramétricos de las densidades $f_X(x)$ y $f(x, y)$. Sean estos estimadores

$$\hat{f}(x, y) = \frac{1}{n h_X h_Y} \sum_{i=1}^n K_X\left(\frac{x - x_i}{h_X}\right) K_Y\left(\frac{y - y_i}{h_Y}\right),$$

$$\hat{f}_X(x) = \frac{1}{n h_X} \sum_{i=1}^n K_X\left(\frac{x - x_i}{h_X}\right) = \int_R \hat{f}(x, y) dy.$$

El estimador de $m(x)$ se obtiene sustituyendo las densidades desconocidas por estos estimadores:

$$\hat{m}(x) = \int_{\mathbb{R}} y \frac{\hat{f}(x, y)}{\hat{f}_X(x)} dy =$$

$$\int_{\mathbb{R}} y \frac{\frac{1}{nh_Xh_Y}\sum_{i=1}^n K_X\left(\frac{x-x_i}{h_X}\right)K_Y\left(\frac{y-y_i}{h_Y}\right)}{\frac{1}{nh_X}\sum_{i=1}^n K_X\left(\frac{x-x_i}{h_X}\right)}dy =$$

$$\frac{\sum_{i=1}^n K_X\left(\frac{x-x_i}{h_X}\right)\int_{\mathbb{R}} y\frac{1}{h_Y}K_Y\left(\frac{y-y_i}{h_Y}\right)dy}{\sum_{i=1}^n K_X\left(\frac{x-x_i}{h_X}\right)}.$$

Haciendo el cambio de variable $u = (y - y_i)/h_Y$ $(y = y_i + h_Y u)$ la integral del numerador es igual a

$$\int_R (y_i + h_Y u)K_Y(u)du = y_i,$$

si hemos usado un núcleo K_Y que integra 1 y cuyo momento de primer orden vale 0. Así, si hacemos $h = h_X$ y $K = K_X$, se tiene que el estimador de $m(x)$ es igual a

$$\hat{m}(x) = \frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)y_i}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)},$$

que es la expresión del estimador de Nadaraya-Watson.

Hay otras formas de asignar los pesos $w(t, x_i)$ en la fórmula genérica

$$\hat{m}_K(t) = \sum_{i=1}^n w(t, x_i)y_i.$$

En particular, cabe mencionar el estimador núcleo de Gasser-Mullër, que sólo difiere del de Nadaraya-Watson en la forma de definir estos pesos. Dada la muestra (x_i, y_i), $i = 1, \ldots, n$, ordenada según el orden creciente de las x_i, se definen los valores s_i así:

$$s_0 = -\infty,\ s_i = \frac{x_i + x_{i+1}}{2}, i = 1, \ldots, n-1,\ s_n = \infty.$$

A partir de estos valores, de un núcleo K y un parámetro de suavizado h, se definen los pesos de Gasser-Müller

$$w(t, x_i)_{GM} = \int_{s_{i-1}}^{s_i} \frac{1}{h}K\left(\frac{u-t}{h}\right)du.$$

El estimador de Gasser-Müller es

$$\hat{m}_{GM}(t) = \sum_{i=1}^n w(t, x_i)_{GM}y_i.$$

El estimador de Nadaraya-Watson es más natural si el diseño es aleatorio, mientras que para el diseño fijo resulta más intuitivo el de Gasser-Müller.

4.2.2. Expresión matricial del estimador por polinomios locales

Se define la matriz

$$X_t = \begin{pmatrix} 1 & (x_1 - t) & \dots & (x_1 - t)^q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n - t) & \dots & (x_n - t)^q \end{pmatrix}.$$

Se definen los vectores $Y = (y_1, \dots, y_n)^{\mathrm{T}}$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^{\mathrm{T}}$, $\beta = (\beta_0, \dots, \beta_q)^{\mathrm{T}}$. Se define la matriz de pesos

$$W_t = \mathrm{Diag}(w(x_1, t), \dots, w(x_n, t)).$$

Ajustamos el modelo

$$Y = X\beta + \varepsilon$$

por mínimos cuadrados generalizados (MCG):

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{q+1}} (Y - X_t\beta)^{\mathrm{T}} W_t (Y - X_t\beta).$$

La solución es

$$\hat{\beta} = \left(X_t^{\mathrm{T}} W_t X_t\right)^{-1} X_t^{\mathrm{T}} W_t Y.$$

Obsérvese que pueden tomarse los pesos

$$w(x_i, t) = \frac{\frac{1}{h} K\left(\frac{x_i - t}{h}\right)}{\frac{1}{h} \sum_{j=1}^{n} K\left(\frac{x_j - t}{h}\right)}$$

o simplemente

$$w(x_i, t) = K\left(\frac{x_i - t}{h}\right)$$

porque en la expresión de $\hat{\beta}$ la matriz W_t aparece elevada a 1 y a (-1), así que se cancelan las constantes multiplicativas que pueda haber en W_t.

Para $j = 0, \dots, q$, sea e_j el vector $(q+1)$-dimensional con todas sus coordenadas iguales a 0, salvo la $(j+1)$-ésima, que vale 1. Entonces

$$\hat{m}_q(t) = \hat{\beta}_0 = e_0^{\mathrm{T}} \hat{\beta} = e_0^{\mathrm{T}} \left(X_t^{\mathrm{T}} W_t X_t\right)^{-1} X_t^{\mathrm{T}} W_t Y = S_t Y,$$

donde $S_t = e_0^{\mathrm{T}} \left(X_t^{\mathrm{T}} W_t X_t\right)^{-1} X_t^{\mathrm{T}} W_t$ es un vector fila n-dimensional. Se tiene entonces que el estimador de la función de regresión por polinomios locales es un ESTIMADOR LINEAL en los datos $y_1, \dots, y_n$.

En general, si se quiere estimar la s-ésima derivada de m en el punto t, se toma

$$\hat{m}_q^{(s)}(t) = s! \hat{\beta}_s(t) = s! e_s^{\mathrm{T}} \hat{\beta},$$

que es también lineal en $y_1, \dots, y_n$.

4.2.3. Propiedades locales de los estimadores por polinomios locales

En el siguiente resultado se dan las expresiones del sesgo y varianza asintóticos del estimador basado en polinomios locales. La demostración puede verse en Fan y Gijbels (1996).

Teorema 4.1 *Se considera el modelo de regresión no paramétrico*

$$Y_i = m(x_i) + \varepsilon_i,\ i = 1 \dots n$$

donde $\varepsilon_1, \dots, \varepsilon_n$ *son v.a. independientes con* $E(\varepsilon_i) = 0$ *y* $V(\varepsilon_i) = \sigma^2(x_i)$*, y el diseño* $x_1, \dots, x_n$ *es regular con función de densidad* $f(x)$*.*

Se suponen las siguientes hipótesis de regularidad:

1. $f(x)$, $m^{(q+2)}(x)$ *y* $\sigma^2(x)$ *son funciones continuas de* x*.*
2. $f(x) > 0$*.*
3. K *es simétrica con soporte en* $[-1, 1]$ *y* $\int_R K(u)du = 1$*.*
4. $(x - h, x + h)$ *está contenido en el soporte de* $f(x)$*.*
5. $h \longrightarrow 0$ *y* $nh \longrightarrow \infty$ *cuando* $n \longrightarrow \infty$*.*

Sea $\hat{m}_q(x)$ *el estimador no paramétrico de* $m(x)$ *basado en el ajuste local de un polinomio de grado* q*. Su varianza asintótica es*

$$V(\hat{m}_q(x)) = \frac{R(K_{(q)})\sigma^2(x)}{nhf(x)} + o\left(\frac{1}{nh}\right).$$

El sesgo asintótico de $\hat{m}_q(x)$ *es como sigue:*

- *Si* q *es impar,* $q = 2k + 1$*,*

$$E(\hat{m}_q(x)) - m(x) = \frac{m^{(q+1)}(x)}{(q+1)!}h^{q+1}\mu_{q+1}(K_{(q+1)}) + o(h^{q+1}) =$$

$$\frac{m^{(2k+2)}(x)}{(2k+2)!}h^{2k+2}\mu_{2k+2}(K_{(2k+2)}) + o(h^{2k+2}).$$

- *Si* q *es par,* $q = 2k$*, (hay que suponer además que* $f'(x)$ *es continua)*

$$E(\hat{m}_q(x)) - m(x) = \left(\frac{m^{(q+1)}(x)f'(x)}{f(x)(q+1)!} + \frac{m^{(q+2)}(x)}{(q+2)!}\right)h^{q+2}\mu_{q+2}(K_{(q+2)}) + o(h^{q+2}) =$$

$$\left(\frac{m^{(2k+1)}(x)f'(x)}{f(x)(2k+1)!} + \frac{m^{(2k+2)}(x)}{(2k+2)!}\right)h^{2k+2}\mu_{2k+2}(K_{(2k+2)}) + o(h^{2k+2}).$$

En estas expresiones, $R(g) = \int_R g(x)^2 dx$, $\mu_j(K) = \int_R u^j K(u)du$, *y* $K_{(j)}$ *es un núcleo de orden* j.

Obsérvese que los grados del polinomio local $q = 2k$ y $q = 2k + 1$ dan resultados similares asintóticamente, puesto que en ambos casos

$$\text{MSE}(\hat{m}_q(x)) = O(h^{4k+4}).$$

En particular el estimador núcleo de Nadaraya-Watson ($q = 0$) y el estimador local lineal ($q = 1$) dan MSE del mismo orden asintótico ($O(h^4)$). Concretamente, para el estimador de Nadaraya-Watson

$$V(\hat{m}_{NW}(x)) = V(\hat{m}_0(x)) = \frac{R(K)\sigma^2(x)}{nhf(x)} + o\left(\frac{1}{nh}\right),$$

$$\text{Sesgo}(\hat{m}_{NW}(x)) = E(\hat{m}_0(x)) - m(x) = \left(\frac{m'(x)f'(x)}{f(x)} + \frac{m''(x)}{2}\right) h^2\mu_2(K) + o(h^2).$$

Para el estimador local lineal,

$$V(\hat{m}_1(x)) = \frac{R(K)\sigma^2(x)}{nhf(x)} + o\left(\frac{1}{nh}\right), \tag{4.3}$$

$$\text{Sesgo}(\hat{m}_1(x)) = E(\hat{m}_1(x)) - m(x) = \frac{m''(x)}{2} h^2\mu_2(K) + o(h^2). \tag{4.4}$$

Se puede probar que el estimador núcleo de Gasser-Müller tiene el siguiente comportamiento en diseño fijo:

$$V(\hat{m}_{GM}(x)) = \frac{R(K)\sigma^2(x)}{nhf(x)} + o\left(\frac{1}{nh}\right),$$

$$\text{Sesgo}(\hat{m}_{GM}(x)) = E(\hat{m}_{GM}(x)) - m(x) = \frac{m''(x)}{2} h^2\mu_2(K) + o(h^2).$$

En diseño aleatorio, sin embargo, el término principal de la varianza del estimador de Gasser-Müller se ve multiplicado por un factor de 3/2, mientras que las varianzas del estimador de Nadaraya-Watson y del estimador local lineal no se alteran.

Es decir, el estimador local lineal aúna las buenas propiedades del estimador núcleo de Nadaraya-Watson (varianzas iguales en diseño fijo y aleatorio) y las del estimador de Gasser-Müller (expresión del sesgo simple y no dependiente de la densidad del diseño).

Las expresiones asintóticas del sesgo son más simples si q es impar (recuerdan las expresiones estudiadas en la estimación núcleo de la densidad)

y no dependen de la densidad $f(x)$ del diseño de las x_i. Es como si los polinomios locales de grado impar se adaptasen al diseño, es decir, a diferentes grados de concentración de la variable explicativa x.

Obsérvese que si el diseño es uniforme ($f(x)$ constante y, por tanto, $f'(x) = 0$) el comportamiento asintótico de los polinomios locales de grados $q = 2k$ y $q = 2k + 1$ es equivalente.

En general se recomienda usar el polinomio local de grado $q = 2k + 1$ en vez de usar el de grado $q = 2k$.

Si se optimiza en h el valor asintótico del MSE de $\hat{m}_q(x)$ se llega a que el valor óptimo de h para estimar $m(x)$ es

$$h_{\text{AMSE}} = O\left(n^{-\frac{1}{4k+5}}\right), \text{ para } q = 2k \text{ o } q = 2k + 1,$$

lo que da lugar a un valor óptimo del AMSE

$$\text{AMSE}^* = O\left(n^{-\frac{4k+4}{4k+5}}\right), \text{ para } q = 2k \text{ o } q = 2k + 1.$$

Por ejemplo, si $q = 0$ o $q = 1$,

$$h_{\text{AMSE}} = O\left(n^{-\frac{1}{5}}\right), \ \text{AMSE}^* = O\left(n^{-\frac{4}{5}}\right),$$

que son los órdenes de convergencia que habíamos visto en estimación de densidades mediante estimadores núcleo con núcleos de orden 2.

Obsérvese que los órdenes de convergencia que se consiguen con polinomios locales de grado $2k$ o $2k + 1$ son análogos a los que se conseguían en estimación de la densidad usando núcleos de alto orden (de orden $2k + 2$, concretamente). En regresión polinómica local se obtienen estos comportamientos tan deseables en el sesgo de los estimadores de una forma más natural que en estimación de la densidad (aumentando el grado del polinomio local, en vez de construyendo núcleos de alto orden).

Si se usan núcleos con soporte compacto y hay zonas de la recta real en las que la densidad del diseño es baja (hay pocos puntos x_i observados) puede ser que $\hat{m}_q(x)$ no esté definida porque no haya ningún dato (x_i, y_i) con $x_i \in [x - h, x + h]$. Una forma de evitar esto es usar un núcleo con soporte en todo $\mathbb{R}$ (núcleo Gaussiano, por ejemplo).

Otra posibilidad es utilizar un estimador núcleo con ventana variable, en la línea de los k vecinos más cercanos. Por ejemplo se puede usar $h(x)$ tal que en $[x-h(x), x+h(x)]$ haya una proporción s dada de puntos x_i, con $s \in (0, 1)$. A s se le suele llamar *span* en los paquetes de estimación no paramétrica. En R, la función `loess` permite hacer estimación polinómica local fijando el valor del `span` s.

4.2.4. Comportamiento en la frontera del soporte de x

Los estimadores núcleo de Nadaraya-Watson o de Gasser-Müller tienen problemas en la frontera del soporte de la variable explicativa, de forma parecida a lo que sucede en la estimación núcleo de la densidad.

En el caso de la regresión el sesgo de los estimadores núcleo es de orden $O(h)$ en la frontera, mientras que en el interior del soporte es de orden $O(h^2)$.

Por su parte, se puede probar que el estimador lineal local corrige automáticamente el problema del sesgo en la frontera y tiene sesgo de orden $O(h^2)$ en todo el soporte de x. Por contra, cerca de la frontera este estimador tiene mayor varianza que el estimador de Nadaraya-Watson.

En general, el estimador polinómico local de grado impar $q = 2k + 1$ tiene menor sesgo en la frontera que el de grado par inmediatamente inferior $q = 2k$. Véase el Ejemplo 4.3 y las dos primeras gráficoas de la Figura 4.4.

4.2.5. Elección del grado del polinomio local

Daremos algunas recomendaciones generales para elegir el grado de los polinomios que se ajustan localmente. Cuanto mayor es q, mejores son las propiedades teóricas del estimador no paramétrico, aunque en la práctica no se aconseja que q supere $(s + 1)$, donde s es el orden de la derivada de $m(x)$ que se desea estimar.

Para estimar la función de regresión, es preferible ajustar polinomios de grado impar a ajustar los del grado par inmediatamente anterior, porque los primeros se adaptan mejor a los datos en la frontera del soporte de la variable explicativa, en el caso de que éste no sea toda la recta real. Por tanto, el estimador lineal local $(q = 1)$ es preferible al estimador de Nadaraya-Watson $(q = 0)$. Señalemos finalmente que el efecto que la elección del grado q tiene en el estimador es mucho menor que el debido a la elección del parámetro de suavizado h.

Para decidir si vale la pena pasar de ajustar un modelo lineal local $(q = 1)$ a ajustar un modelo cúbico local $(q = 3)$, hay que tener en cuenta la expresión asintótica del sesgo del estimador local lineal:

$$\text{Sesgo}(\hat{m}_1(x)) = \frac{m''(x)}{2} h^2 \mu_2(K) + o(h^2).$$

Obsérvese que el sesgo será alto en zonas donde la función $m(x)$ tenga gran curvatura ($|m''(x)|$ grande). Por tanto, si $m(x)$ presenta cambios abruptos conviene usar $q = 3$ en vez de $q = 1$.

Ejemplo 4.3

La Figura 4.4 muestra diversos ajustes por polinomios locales de la función de regresión de `RM` sobre `LSTAT` (conjunto de datos de características de las viviendas en los barrios de Boston). En todos se ha usado el núcleo de Epanechnikov y ventana $h = 7$.

Se observa que el ajuste en la frontera del soporte de la variable explicativa mejora al aumentar el grado del polinomio local (la mejora es más notable al pasar de $q = 0$ a $q = 1$).

También se aprecia que para $q = 2$ y $q = 3$ hay menos sesgo que para $q = 0$ o $q = 1$ en las zonas en las que la función de regresión tiene más curvatura (con $q = 2$ y $q = 3$ se aprecia la presencia de un mínimo y un máximo local cerca de `LSTAT` igual a 15 y a 20, respectivamente).

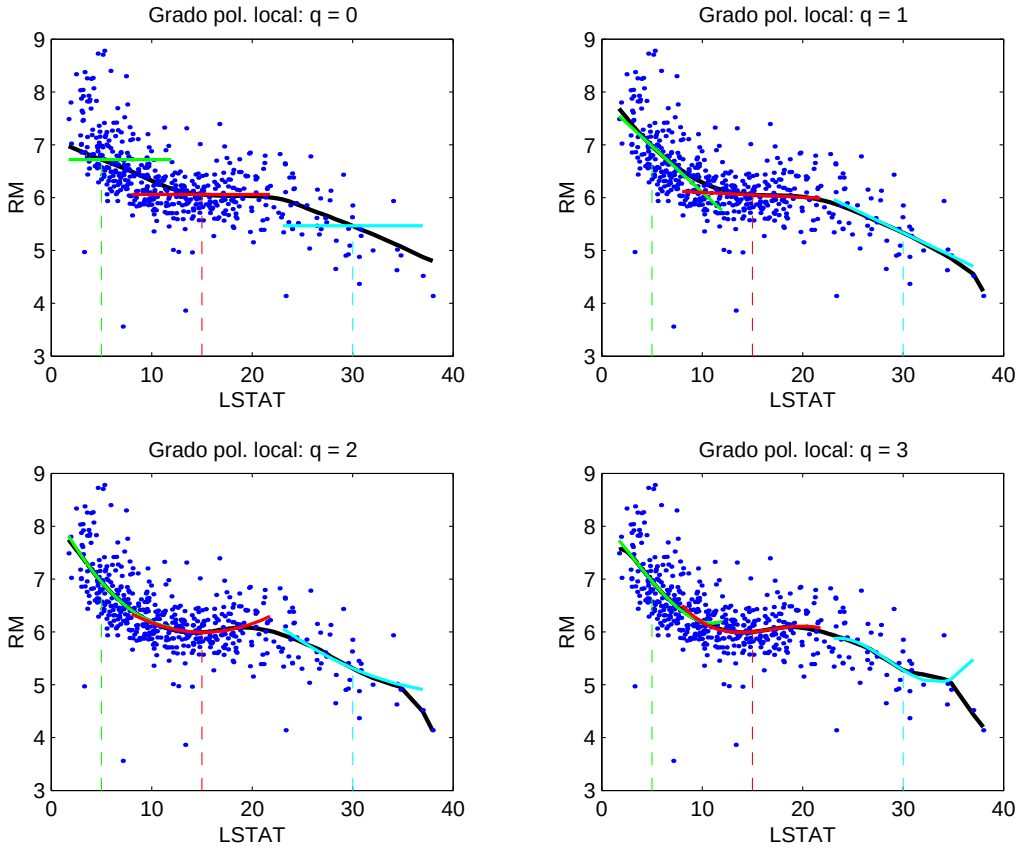


Figura 4.4: Ajustes polinómicos locales de la variable `RM` como función de `LSTAT`. Se han usado grados $q = 0, 1, 2, 3$, de los polinomios locales.

Ejemplo 4.4

La Figura 4.5 muestra la estimación de la derivada de la función de regresión de `RM` sobre `LSTAT`. Se han ajustado polinomios locales de segundo grado. Se ha usado el núcleo de Epanechnikov y ventana $h = 7$.

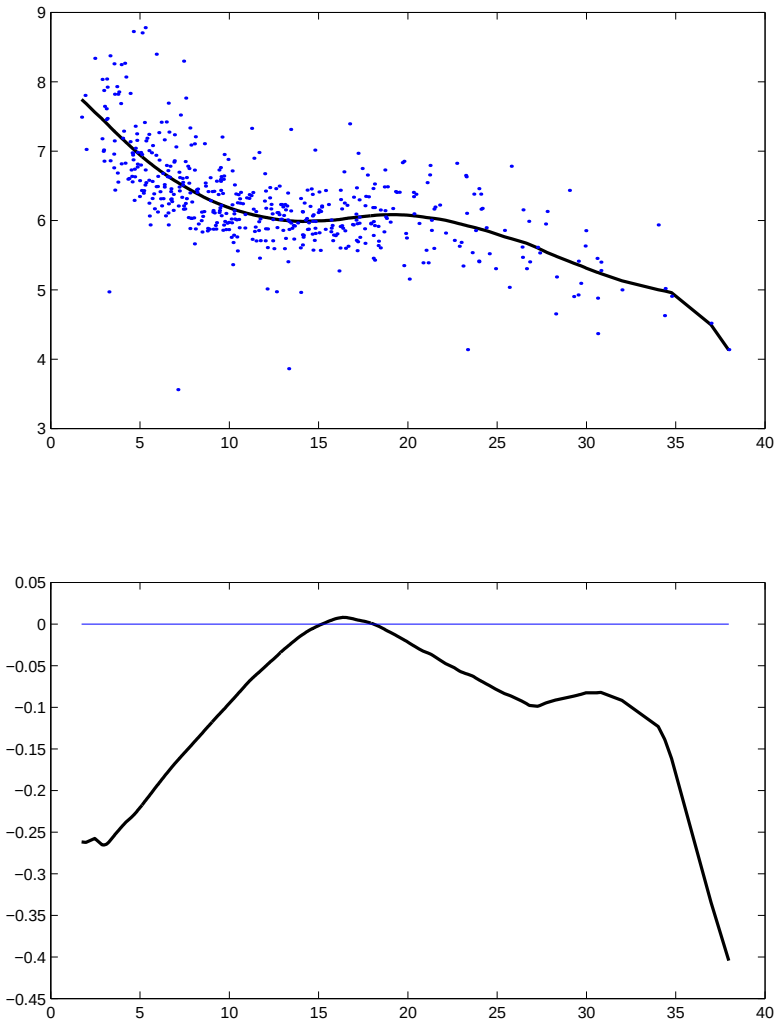


Figura 4.5: Estimación de la derivada de la función de regresión de `RM` sobre `LSTAT` Se ha usado un polinomio local de grado 2.

4.3. Elección del parámetro de suavizado

Como se ha mencionado anteriormente, la elección del parámetro de suavizado h tiene una importancia crucial en el aspecto y propiedades del estimador de la función de regresión. En la práctica, valores distintos de h pueden producir estimadores completamente distintos. La Figura 4.6 muestra tres estimaciones de la función de regresión correspondientes a otros tantos valores del parámetro de suavizado: $h = 0{,}25$ (excesivamente pequeño: el estimador es muy poco suave y tiene muchas irregularidades), $h = 2{,}5$ (es el que se usó en la Figura 3; parece una buena elección) y $h = 15$ (demasiado grande: se suaviza demasiado y el estimador no paramétrico es casi igual al paramétrico, la recta de regresión).

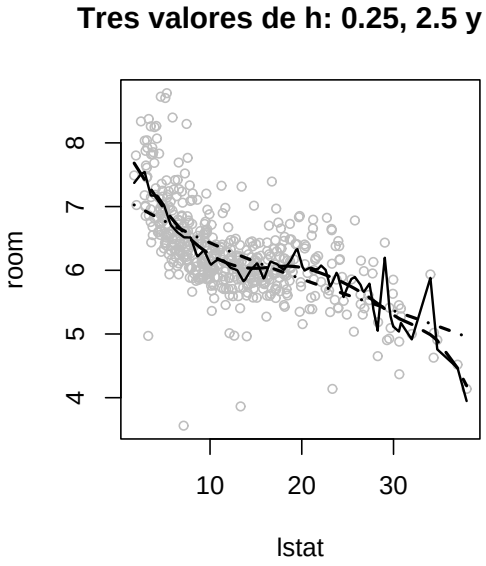


Figura 4.6: Ajuste lineal local con núcleo gaussiano y tres valores del parámetro de suavizado h.

El parámetro de suavizado controla el equilibrio que el estimador no paramétrico de la función de regresión debe mantener entre el buen ajuste a los datos observados y la capacidad de predecir bien observaciones futuras. Valores pequeños de h dan mucha flexibilidad al estimador y le permiten acercarse a todos los datos observados (cuando h tiende a 0 el estimador acaba por interpolar los datos), pero los errores de predicción asociados serán altos. Hay, por tanto, sobreajuste (*overfitting*). En el caso de que h tome un

tamaño moderado no se ajustará tan bien a las observaciones (tampoco es necesario, dado que los datos pueden contener ruido aleatorio) pero predecirá mejor. En el otro extremo, si h es demasiado grande, tendremos falta de ajuste (*underfitting*), como puede ocurrir con los modelos paramétricos globales.

Buscar el valor adecuado del parámetro de suavizado persigue conseguir el equilibrio entre el *sesgo* y la *varianza* del estimador. Para h pequeño el estimador es muy variable (aplicado a muestras distintas provenientes del mismo modelo da resultados muy distintos) y tiene poco sesgo (el promedio de los estimadores obtenidos para muestras distintas es aproximadamente la verdadera función de regresión). Si h es grande ocurre lo contrario.

El parámetro de suavizado puede elegirse de forma *manual*: comparando los resultados obtenidos para distintos valores de h y eligiendo aquél que, a juicio del investigador, dé el resultado más satisfactorio visualmente, o el más informativo (el que mejor resuma la relación existente entre los datos). Esta forma de proceder está sujeta a la opinión subjetiva del usuario y no puede automatizarse, lo que la hace inviable cuando el número de estimaciones no paramétricas que se han de realizar es grande. Se necesitan, pues, métodos automáticos de selección del parámetro de suavizado. Citaremos aquí los más habituales.

4.3.1. Error de predicción en una muestra test

Éste es un método que suele usarse en los campos del aprendizaje automático y la minería de datos para calibrar métodos de predicción. Si la cantidad de datos disponibles permite dividir éstos en una muestra para la estimación del modelo (conjunto de entrenamiento) y una muestra test, entonces una buena medida de la calidad de un valor h del parámetro de suavizado es el error cuadrático medio de predicción en la muestra test:

$$\mathrm{ECMP}_{\mathrm{test}}(h) = \frac{1}{n_t}\sum_{i=1}^{n_t}(y_i^{\mathrm{test}} - \hat{m}(x_i^{\mathrm{test}}))^2,$$

donde $(x_i^{\mathrm{test}}, y_i^{\mathrm{test}})$, $i = 1, \ldots, n_t$, es la muestra test y $\hat{m}(x)$ es el estimador no paramétrico construido con parámetro de suavizado h usando la otra parte de la muestra original (la de entrenamiento). Se elige como parámetro de suavizado el valor h_{test} que minimiza esa función.

4.3.2. Validación cruzada

Es una técnica usada en muchos campos para la elección de parámetros que controlan el equilibrio entre precisión y variabilidad (o entre bondad

del ajuste y capacidad predictiva) cuando no hay posibilidad de disponer de una muestra test. Consiste en sacar de la muestra consecutivamente cada una de las observaciones x_i, estimar el modelo con los restantes datos (sea $\hat{m}_{(i)}(x)$ el estimador así obtenido), predecir el dato ausente con ese estimador (así se está haciendo de hecho predicción fuera de la muestra) y, finalmente, comparar esa predicción con el dato real. Esto se hace con cada posible valor de h, lo que permite construir la función

$$\mathrm{ECMP}_{\mathrm{CV}}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}_{(i)}(x_i))^2,$$

que mide el error de predicción del estimador fuera de la muestra para cada h. El valor que minimice esa función, h_{CV}, será el valor del parámetro de suavizado elegido.

Ejemplo 4.5

La Figura 4.7 muestra el gráfico de la función $\mathrm{ECMP}_{\mathrm{CV}}(h)$ en el ejemplo que venimos usando (relación entre las variable `RM` y `LSTAT`). Se han usado polinomios de grado 1 con pesos dados por un núcleo gaussiano.

Se observa que tanto los valores de h excesivamente pequeños como los excesivamente grandes dan lugar a errores de predicción fuera de la muestra excesivamente grandes, y que el óptimo se encuentra en un valor intermedio, $h_{\mathrm{CV}} = 2{,}12$, que dista poco del valor $h = 2{,}5$ usado en el panel derecho de la Figura 4.3.

.

4.3.3. Validación cruzada generalizada.

En los estimadores de la función de regresión que, como el basado en ajuste de polinomios locales, son lineales en las observaciones de la variable dependiente se puede probar que para calcular el $\mathrm{ECMP}_{\mathrm{CV}}$ (error cuadrático medio de predicción) no es necesario ajustar las n regresiones que se tienen dejando fuera de la muestra cada uno de los n casos observados (x_i, y_i).

En los estimadores lineales la predicción de la función de regresión en cada valor observado x_i es

$$\hat{y}_i = \sum_{j=1}^{n} w^*(x_i, x_j)y_j.$$

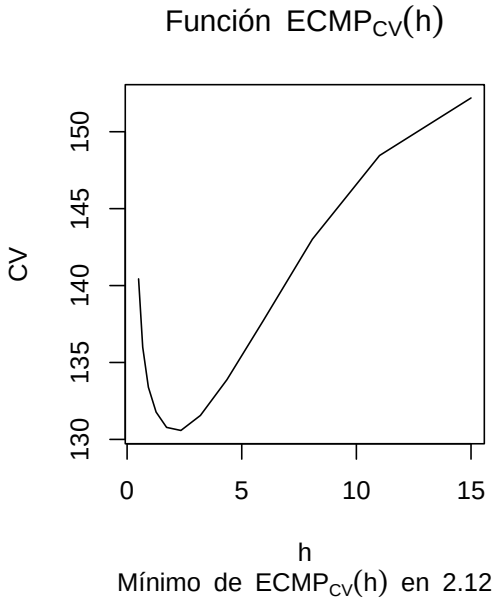


Figura 4.7: Valores del error (validación cruzada) según el parámetro de suavizado h, en la estimación lineal local de `RM` como función de `LSTAT`.

En forma matricial tenemos que

$$\hat{Y} = SY,$$

donde los vectores columna Y e $\hat{Y}$ tienen elementos y_i e $\hat{y}$, respectivamente, y la matriz S (llamada *matriz de suavizado*) tiene su elemento (i, j), s_{ij}, igual a $w^*(x_i, x_j)$. La matriz de suavizado es análoga a la matriz sombrero $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ en regresión lineal múltiple:

$$\hat{Y}_L = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y = HY.$$

Se puede demostrar que

$$\mathrm{ECMP}_{\mathrm{CV}}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - s_{ii}}\right)^2,$$

(igual que ocurre en el modelo de regresión lineal múltiple), con lo que para evaluar esta función no es necesario ajustar n regresiones no paramétricas, sino que basta con ajustar únicamente la que hace intervenir todos los datos y anotar la matriz S.

En estos modelos puede hacerse una modificación del criterio de la validación cruzada que se denomina VALIDACIÓN CRUZADA GENERALIZADA. Esta modificación consiste en sustituir en la fórmula anterior los valores s_{ii} de la diagonal de S por su valor promedio:

$$\mathrm{ECMP}_{\mathrm{GCV}}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-\nu/n}\right)^2,$$

donde

$$\nu = \mathrm{Traza}(S) = \sum_{i=1}^{n} s_{ii}$$

es la suma de los elementos de la diagonal de S.

En el caso de la regresión lineal múltiple ($S = H$) con k regresores (X es una matriz $n \times k$) incluyendo el término independiente, se tiene que

$$\mathrm{Traza}(H) = \mathrm{Traza}(X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}) =$$

$$\mathrm{Traza}((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X) = \mathrm{Traza}(I_k) = k,$$

que es el número de parámetros del modelo. Por eso a $\nu = \mathrm{Traza}(S)$ se le llama NÚMERO DE PARÁMETROS EFECTIVOS del estimador no paramétrico correspondiente a la matriz S.

Tras manipular esta expresión se tiene que

$$\mathrm{ECMP}_{\mathrm{GCV}}(h) = \frac{n\hat{\sigma}^2_{\varepsilon}}{n-\nu},$$

donde

$$\hat{\sigma}^2_{\varepsilon} = \frac{1}{n-\nu}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.5}$$

es un estimador de la varianza de los errores del modelo.

El valor h_{GCV} que minimiza esa función es el parámetro de suavizado elegido por el criterio de validación cruzada generalizada.

En general $\nu = \nu(h)$ es decreciente en h. Dado que la interpretación de ν, como número de parámetros efectivos, es la misma sea cual sea el estimador (siempre que sea lineal en el sentido expresado más arriba) algunos paquetes estadísticos admiten que el grado de suavidad de la estimación no paramétrica se exprese en términos de ν, en vez de hacerlo en términos de h.

4.3.4. Plug-in

Vamos a centrarnos ahora en el estimador lineal local, $\hat{m}(x)$. Supondremos que la densidad del diseño $f(x)$ tiene soporte $[a, b] \subseteq \mathbb{R}$. También supondremos que el modelo de regresión no paramétrico es homocedástico ($V(Y|X = x) = \sigma^2(x) = \sigma^2$). Presentamos la propuesta de Ruppert, Sheather y Wand (1995) para elegir la ventana h por el método *plug-in*.

Una medida global del ajuste del estimador $\hat{m}$ a la verdadera función m es el error cuadrático medio integrado:

$$\text{MISE}(\hat{m}) = E_{\mathbf{Z}}(\text{ISE}(\hat{m})) = E_{\mathbf{Z}}\left(\int_a^b (\hat{m}(x) - m(x))^2 f(x)dx\right).$$

En esta ecuación $\mathbf{Z}$ representa la muestra aleatoria de tamaño n a partir de la cual se construye el estimador no paramétrico $\hat{m}$: $\mathbf{Z} = \{(x_i, y_i) : i = 1, \ldots, n\}$.

Al igual que ocurría en estimación de la densidad, se tiene que $\text{MISE}(\hat{m}) = \text{IMSE}(\hat{m})$. Teniendo en cuenta las expresiones (4.3) y (4.4) de la varianza y el sesgo asintóticos del estimador local lineal, se tiene que

$$\text{MISE}(\hat{m}) = \text{IMSE}(\hat{m}) = \frac{h^4\mu_2^2(K)}{4}\int_a^b (m''(x))^2 f(x)dx + \frac{R(K)\sigma^2}{nh} + o\left(h^4 + \frac{1}{nh}\right).$$

Así, el valor de h que minimiza el AMISE (la parte principal del MISE) es

$$h_0 = \left(\frac{R(K)\sigma^2}{\mu_2^2(K)\int_a^b (m''(x))^2 f(x)dx}\right)^{1/5} n^{-1/5}.$$

El método de selección de h por PLUG-IN, que lleva a seleccionar una ventana que denotaremos por h_{PI}, consiste en sustituir en esta expresión las cantidades desconocidas por estimaciones de ellas. En concreto, para dar un valor a h_0 necesitamos:

(i) estimar $\int_a^b (m''(x))^2 f(x)dx$,

(ii) estimar $\sigma^2 = V(Y|X = x) = V(\varepsilon)$.

Estimar el valor esperado de $(m''(X))^2$:

Para estimar

$$\int_a^b (m''(x))^2 f(x)dx = E[(m''(X))^2],$$

donde $X \sim f(x)$, se puede seguir el siguiente procedimiento. Se ajusta a los datos (x_i, y_i), $i = 1, \ldots, n$, un estimador polinómico local de tercer grado con pesos dados por el núcleo K y una ventana g que habrá que determinar: $w(x_i, t) = K((x_i - t)/g)$. La estimación de la segunda de derivada de m en un punto t se calcula como vimos en la ecuación (4.2).

De esta forma se estima $m''(t)$ para $t = x_1, \ldots, x_n$. Entonces $E[(m''(X))^2]$ se estima como

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{m}_g''(x_i))^2.$$

El valor de g óptimo para estimar la segunda derivada de $m(x)$ es

$$g_0 = C_2(K)\left(\frac{\sigma^2}{|\int_a^b m''(x)m^{(iv)}(x)f(x)dx|}\right)^{1/7} n^{-1/7}.$$

La estimación de $m''(x)$ y $m^{(iv)}(x)$ se hace dividiendo el rango de los datos x_i en subintervalos y ajustando un polinomio de grado 4 (o mayor) en cada subintervalo. Este método también da una primera estimación de σ^2.

Estimar el valor de σ^2:

Hay varias opciones para estimar la varianza de los errores en el modelo de regresión no paramétrica.

1. Acabamos de ver una de ellas: dividir el rango de la variable explicativa en subintervalos y ajustar modelos paramétricos en cada uno de ellos.

2. La ecuación (4.5) ofrece otra forma de estimar $\sigma^2 = V(\varepsilon)$. Allí se usa la suma de cuadrados de los errores en la muestra original divididos por n menos el número de parámetros efectivos del modelo no paramétrico ν. Esa ecuación se sigue de la analogía entre el estimador de la regresión lineal múltiple y los estimadores no paramétricos lineales.

3. Esta analogía puede sugerir aún otro estimador de σ^2. En el modelo de regresión lineal múltiple con k regresores

$$Y = X\beta + \varepsilon,\ \hat{Y}_L = HY,\ \hat{\varepsilon} = Y - \hat{Y}_L = (I - H)Y.$$

Si $\varepsilon \sim N(0, \sigma^2 I)$ se tiene que

$$\frac{1}{\sigma^2}\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon} = \varepsilon^{\mathrm{T}}(I - H)^{\mathrm{T}}(I - H)\varepsilon \sim \chi^2_{n-k}.$$

A la cantidad $(n-k)$ se le llama GRADOS DE LIBERTAD del modelo, y esa cantidad es la traza de la matriz que define la forma cuadrática en ε:

$$\text{Traza}((I-H)^{\mathrm{T}}(I-H)) = \text{Traza}((I-H)(I-H)) =$$

$$\text{Traza}(I-H) = \text{Traza}(I) - \text{Traza}(H) = n-k.$$

Obsérvese que se ha utilizado que la matriz H es simétrica e idempotente: $H = H^{\mathrm{T}}$ y $H^2 = H$. Lo mismo le ocurre a $(I-H)$.

Como la esperanza de una χ^2 es igual a su número de grados de libertad, se tiene que

$$\hat{\sigma}^2 = \frac{1}{n-k}\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon} = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

es un estimador insesgado de σ^2.

En el caso de un estimador lineal de la regresión no paramétrica, la matriz de suavizado S hace el papel de la matriz sombrero H: $\hat{Y} = SY$. Así, el número de GRADOS DE LIBERTAD EFECTIVOS del modelo no paramétrico es

$$\eta = \text{Traza}((I-S)^{\mathrm{T}}(I-S)) = \text{Traza}(I - S^{\mathrm{T}} - S + S^{\mathrm{T}}S) = \tag{4.6}$$

$$n - 2\text{Traza}(S) + \text{Traza}(S^{\mathrm{T}}S).$$

Obsérvese que en este caso, la matriz S no es ni simétrica ni idempotente y por lo tanto $\text{Traza}(S) \neq \text{Traza}(S^{\mathrm{T}}S)$.

Ya se definió $\nu = \text{Traza}(S)$, como número de parámetros efectivos del modelo. Se puede definir $\tilde{\nu} = \text{Traza}(S^{\mathrm{T}}S)$. Así, el número de grados de libertad efectivos es

$$\eta = n - 2\nu + \tilde{\nu}.$$

Se define entonces el estimador de σ^2 como

$$\hat{\sigma}^2 = \frac{1}{n-2\nu+\tilde{\nu}}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Con esta definición $\hat{\sigma}^2$ es insesgado si la función de regresión $m(x)$ es lineal. También se puede probar que si $m(x)$ es suficientemente suave entonces $\hat{\sigma}^2$ es un estimador consistente de σ^2.

Esta estimación de σ^2 requiere que se haya estimado de alguna manera el modelo de regresión no paramétrica. Si se ajusta un estimador local lineal con núcleo K, se puede probar que la ventana óptima para

estimar σ^2 es

$$C_3(K)\left(\frac{\sigma^4}{(\int_a^b (m''(x))^2 f(x)dx)^2}\right) n^{-2/9}.$$

En esta expresión, σ^2 y m'' se estiman como se indicó más arriba: dividiendo el rango de las x_i en subintervalos y ajustando modelos paramétricos en cada uno de ellos.

4. Propuesta de Rice (1984).
Consideramos que en el modelo de regresión $y_i = m(x_i) + \varepsilon_i$, $i = 1, \ldots, n$, los datos están ordenados según el orden creciente de x_i. Así,

$$y_i - y_{i-1} = m(x_i) - m(x_{i-1}) + (\varepsilon_i - \varepsilon_{i-1}).$$

Por tanto,

$$V(y_i - y_{i-1}) = E\left[(y_i - y_{i-1})^2\right] - (m(x_i) - m(x_{i-1}))^2 = V\left[(\varepsilon_i - \varepsilon_{i-1})^2\right] = 2\sigma^2.$$

Si la función m es suficientemente suave y los puntos x_i y x_{i-1} son suficientemente próximos, la cantidad $(m(x_i) - m(x_{i-1}))^2$ es despreciable comparada con $E\left[(y_i - y_{i-1})^2\right]$, y se tiene que

$$E\left[(y_i - y_{i-1})^2\right] \approx 2\sigma^2,$$

de donde se sigue que

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

es un estimador aproximadamente insesgado de σ^2.

5. Propuesta de Gasser, Sroka y Jennen-Steinmetz (1986)
Se basa en realizar interpolaciones lineales alrededor de cada observación (x_i, y_i), usando para ello las observaciones (x_{i-1}, y_{i-1}) y (x_{i+1}, y_{i+1}) (se supone que los datos están ordenados según x_i). Sea

$$\hat{y}_i = \frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}} y_{i-1} + \frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}} y_{i+1} = a_i y_{i-1} + b_i y_{i+1}$$

el interpolador lineal de (x_{i-1}, y_{i-1}) y (x_{i+1}, y_{i+1}) evaluado en $x = x_i$.

Se define

$$\tilde{\varepsilon}_i = \hat{y}_i - y_i = a_i y_{i-1} + b_i y_{i+1} - y_i.$$

Su esperanza es

$$E(\tilde{\varepsilon}_i) = a_i m(x_{i-1}) + b_i m(x_{i+1}) - m(x_i) = \hat{m}_l(x_i) - m(x_i) \approx 0,$$

donde $\hat{m}_l(x_i)$ es el interpolador lineal de $(x_{i-1}, m(x_{i-1}))$ y $(x_{i+1}, m(x_{i+1}))$ evaluado en $x = x_i$, que es aproximadamente igual a $m(x_i)$ si la función m es suficientemente suave y los puntos x_{i-1} y x_{i+1} son suficientemente próximos.

Así,

$$E(\tilde{\varepsilon}_i^2) \approx V(\tilde{\varepsilon}_i) = (a_i^2 + b_i^2 + 1)\sigma^2,$$

lo que implica que

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=2}^{n-1} \frac{1}{a_i^2 + b_i^2 + 1} \tilde{\varepsilon}_i^2$$

es un estimador aproximadamente insesgado de σ^2.

4.3.5. Comportamiento asintótico de selectores de h

Hemos visto tres selectores del parámetro de suavizado h que no requieren de una muestra test: h_{CV}, h_{GCV} y h_{PI}. Estos tres selectores convergen al valor h_0 que minimiza el AMISE cuando n tiende a infinito, pero lo hacen a diferentes velocidades:

$$\frac{h_{\mathrm{CV}}}{h_0} - 1 = O_p(n^{-1/10}),\ \frac{h_{\mathrm{GCV}}}{h_0} - 1 = O_p(n^{-1/10}),\ \frac{h_{\mathrm{PI}}}{h_0} - 1 = O_p(n^{-2/7}).$$

4.3.6. Ventana variable

La expresión de la ventana h_{AMSE} que minimiza el error cuadrático medio asintótico puntual, AMSE, del estimador lineal local $\hat{m}(t)$ como estimador de $m(t)$ es (ver ecuaciones 4.3 y 4.4)

$$h_{\mathrm{AMSE}}(t) = \left(\frac{R(K)\sigma^2(t)}{\mu_2^2(K) f(t) (m''(t))^2} \right)^{1/5} n^{-1/5}.$$

Esta expresión sugiere que en ocasiones es preferible usar parámetros de suavizado que dependan del punto t donde se está estimando la función de regresión ($h(t)$) o de los valores observados de la variable explicativa ($h(x_i)$):

- Cuando la densidad de la variable explicativa varíe considerablemente a lo largo de su recorrido (en zonas con muchos datos la ventana puede ser más pequeña que en zonas donde haya pocas observaciones).

  En el ejemplo que hemos seguido durante este capítulo de la regresión de la variable `RM` sobre la variable `LSTAT` vemos que la densidad de la variable explicativa es mucho menor con valores altos de `LSTAT`.

- Cuando la varianza de los errores sea función de la variable explicativa (en zonas con gran variabilidad en los errores es recomendable usar valores grandes de la ventana).

- cuando la curvatura de la función de regresión sea diferente en diferentes tramos del recorrido de la variable explicativa (en zonas donde la variabilidad sea grande se deben usar valores más pequeños de h).

La forma más habitual de incluir una ventana variable en el estimador no paramétrico es fijar la proporción s de puntos que se desea usar en la estimación de cada valor $m(t)$ y definir $h(t)$ tal que el número de datos (x_i, y_i) con x_i perteneciente al intervalo $(t - h(t), t + h(t))$ sea sn. La proporción s se denomina *span*.

Si se ajusta un polinomio de grado $q = 0$ (estimador de Nadaraya-Watson), se usa el núcleo uniforme y se elige $s = k/n$, el estimador resultante es el estimador de los k vecinos más cercanos (*k-nearest neighbours*, en inglés). La elección de s (o de $k = sn$) puede hacerse mediante validación cruzada o usando una muestra test.

4.4. Verosimilitud local

4.4.1. Discriminación no paramétrica mediante regresión binaria local

La discriminación no paramétrica basada en la estimación de las funciones de densidad en cada subpoblación (ver la Sección 3.6.7) no es la única vía de plantearse el problema de clasificación desde una óptica no paramétrica. Este problema puede también modelarse como uno de regresión no paramétrica en el que la respuesta (el indicador de la clase a la que pertenece cada dato) es categórica. Aquí nos limitaremos a estudiar el caso en el que sólo hay dos clases (respuesta binaria). Estos modelos son una versión no paramétrica de los modelos lineales generalizados.

Consideremos el problema de análisis discriminante con dos clases, C_0 y C_1, y observaciones asociadas

$$(y_i; x_{i1}, \ldots, x_{ip}),\ i = 1, \ldots, n,$$

con y_i igual a 0 o a 1, según si la observación i-ésima pertenece a una u otra clase.

Si la variable explicativa X es unidimensional, su densidad marginal es la mixtura

$$f(x) = \alpha_0 f(x|C_0) + \alpha_1 f(x|C_1) = \alpha_0 f(x|Y = 0) + \alpha_1 f(x|Y = 1),$$

donde $\alpha_i = P(C_i) = P(Y = i)$, $i = 1, 2$, son las probabilidades a priori de cada clase.

Si dx es una longitud suficientemente pequeña,

$$P(X \in [x, x + dx]) \approx f(x)|dx| = \alpha_0 f(x|Y = 0)|dx| + \alpha_1 f(x|Y = 1)|dx|.$$

Por el Teorema de Bayes, para $i = 0, 1$ se tiene que

$$P(Y = i|X \in [x, x + dx]) \approx \frac{\alpha_i f(x|Y = i)|dx|}{\alpha_0 f(x|Y = 0)|dx| + \alpha_1 f(x|Y = 1)|dx|} = \frac{P(Y = i) f(x|Y = i)}{f(x)}.$$

Concluimos que

$$m(x) = E(Y|X = x) \approx P(Y = 1|X \in [x, x + dx]) \approx \frac{P(Y = 1) f(x|Y = 1)}{f(x)}.$$

Se puede probar que en efecto se da la igualdad

$$m(x) = E(Y|X = x) = P(Y = 1) f(x|Y = 1)/f(x).$$

El mismo tipo de razonamiento se puede hacer para una variable explicativa X p-dimensional (el incremento dx también es p-dimensional y en las expresiones anteriores $|dx|$ debe leerse como volumen del hipercubo de lados iguales a las componentes de dx).

Dados los datos

$$(y_i; x_{i1}, \ldots, x_{ip}),\ i = 1, \ldots, n,$$

modelizamos y_i como una variable aleatoria que toma los valores 1 y 0 con probabilidades respectivas p_i y $1 - p_i$, donde p_i es función de las variables explicativas $(x_{i1}, \ldots, x_{ip})$:

$$p_i = E(Y_i|x_{i1}, \ldots, x_{ip}) = P(Y_i = 1|x_{i1}, \ldots, x_{ip}) =$$

$$m(x_{i1},\ldots,x_{ip}) = P(Y=1)f(x_{i1},\ldots,x_{ip}|Y=1)/f(x_{i1},\ldots,x_{ip}).$$

La distribución de probabilidad queda entonces totalmente determinada: y_i sigue una Bernoulli de parámetro $p_i = m(x_{i1},\ldots,x_{ip})$.

Si la función $m(x_{i1},\ldots,x_{ip})$ fuese conocida, tendríamos una forma sencilla de clasificar una nueva observaciones de la que sólo conociésemos las variables explicativas $x = (x_1,\ldots,x_p)$: se clasificaría en la población C_1 si y sólo si

$$p_x = m(x_1,\ldots,x_p) > 0{,}5 \Longleftrightarrow$$

$$P(Y=1)f(x_1,\ldots,x_p|Y=1) > P(Y=0)f(x_1,\ldots,x_p|Y=0),$$

que es la regla Bayes.

Dado que en general la función m no será conocida, lo que se propone es sustituirla en esa regla de clasificación por una estimación suya hecha de forma no paramétrica.

La estimación de $p_x = m(x_1,\ldots,x_p)$ sería fácil si un modelo paramétrico (digamos el modelo logístico) se adaptase bien a todo el rango de valores de las variables explicativas. Esto no siempre es así y por eso precisamente buscamos un estimador no paramétrico. No obstante, aunque globalmente el modelo logístico no sea una buena aproximación de la función, sí lo puede ser localmente, en un entorno del punto $x = (x_1,\ldots,x_p)$. En realidad, es el mismo tipo de aproximación que hacíamos al estimar localmente mediante un polinomio la función de regresión no paramétrica.

Por lo tanto, suponemos que si $x_i = (x_{i1},\ldots,x_{ip})$ está en un entorno de $x = (x_1,\ldots,x_p)$ entonces y_i sigue un modelo logístico:

$$p_i = \frac{1}{1+e^{-\beta^{\mathrm{T}}x_i}}, \text{ o de forma equivalente } \log\left(\frac{p_i}{1-p_i}\right) = \beta^{\mathrm{T}}x_i.$$

Obsérvese que el vector de parámetros β es función del punto $x = (x_1,\ldots,x_p)$, porque ese punto es el que define qué observaciones están en su entorno y cuáles no.

Resulta pues que en un entorno de $x = (x_1,\ldots,x_p)$ ajustamos un modelo paramétrico que podemos estimar, por ejemplo, por máxima verosimilitud. La contribución de cada observación a la función de log-verosimilitud es, por tratarse de un modelo logístico,

$$y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i).$$

Sumando sobre todas los observaciones y ponderando cada una por un peso $w_i = w(x,x_i)$ (decreciente en la distancia que separa x_i de x) se obtiene la

llamada *función de log-verosimilitud local*:

$$l_x(\beta) = \sum_{i=1}^{n} w_i \left(y_i \log \left(\frac{p_i}{1-p_i} \right) + \log(1-p_i) \right).$$

Maximizando esta función se obtiene un estimador de β, $\hat{\beta}$, que permite obtener una estimación de $p_x = m(x_1, \ldots, x_p)$:

$$\hat{m}(x_1, \ldots, x_p) = \hat{p}_x = \frac{1}{1 + e^{-\hat{\beta}^{\mathrm{T}} x}}.$$

Según si este valor es menor o mayor que 0.5, se clasificará la observación $x = (x_1, \ldots, x_p)$ en C_0 o en C_1, respectivamente.

En la práctica los pesos $w_i = w(x, x_i)$ se definen a partir de funciones núcleo en dimensión p, del mismo modo que se hace en estimación no paramétrica de la densidad multivariante.

El modelo logístico, elegido como aproximación paramétrica local, puede cambiarse por otro modelo paramétrico de respuesta binaria. La variación en los resultados es poco perceptible, porque la función de verosimilitud local usa únicamente pequeñas partes del modelo paramétrico elegido, y dos modelos paramétricos distintos pueden tener partes pequeñas semejantes, aunque globalmente sean distintos.

Ejemplo 4.6

La Figura 4.8 muestra la puesta en práctica de la regresión binaria no paramétrica mediante el ajuste local del modelo logístico. En el conjunto de datos sobre características de las viviendas en barrios de Boston, se desea recuperar la variable binaria creada a partir de `RM` ($y_i = 0$ si `RM`$<$ 6,2, $y_i = 1$ en caso contrario) como función de la variable `LSTAT`. En cada ajuste local se han definido los pesos de cada observación según un núcleo gaussiano con ventana (desviación típica) igual a 3. El punto en el que la función de probabilidad estimada cruza el valor 0.5 es `LSTAT`$=$ 10. Por lo tanto, la regla de regresión logística no paramétrica predice $y = 0$ cuando el valor observado de `LSTAT` es mayor que 10, y predice $y = 1$ en caso contrario. Este resultado es muy similar al obtenido con la regla discriminante vista en la sección anterior (9.38).

Completamos el ejemplo incluyendo una segunda variable explicativa (`AGE`: porcentaje de viviendas construidas antes de 1940 en cada barrio de Boston) en el modelo de regresión logística local. La Figura 4.9 muestra las curvas de nivel de la estimación no paramétrica de la probabilidad de pertenecer a la clase C_2 en función de (`LSTAT`, `AGE`). La línea de trazo grueso

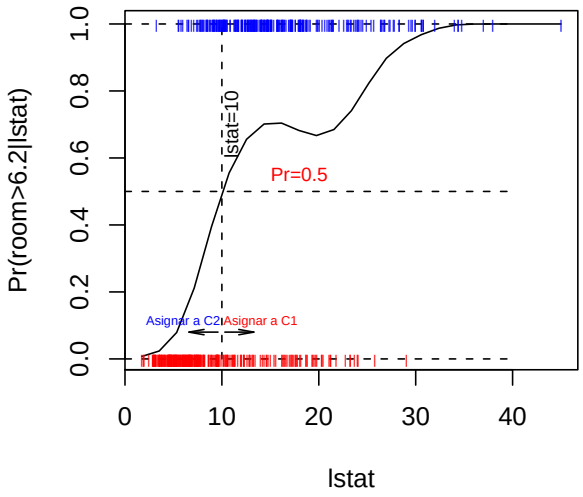


Figura 4.8: Regla discriminante basada en regresión logística local.

está formada por aquellos puntos en los que la estimación de esta probabilidad es igual a 0.5. Por tanto esta línea es la frontera de las zonas que se clasificarán en C_1 (a la izquierda de la línea) o en C_2 (a la derecha). Si se compara esta figura con la Figura 14 se aprecian grandes similitudes, aunque no hay una coincidencia total.

4.4.2. Modelo de verosimilitud local

En esta sección veremos que el modelo de regresión no paramétrica, estimado mediante regresión lineal local, y el modelo de regresión binaria no paramétrica, estimado mediante regresión logística local, son dos casos particulares de lo que podríamos llamar modelo de regresión no paramétrica (general) estimado mediante el ajuste de modelos paramétricos mediante verosimilitud local.

Consideremos el modelo de regresión 4.1 presentado al inicio de este capítulo:

$$y_i = m(x_i) + \varepsilon_i,\ i = 1, \ldots, n,$$

y supongamos que los residuos son independientes y normales: $\varepsilon_i \sim N(0, \sigma^2)$.

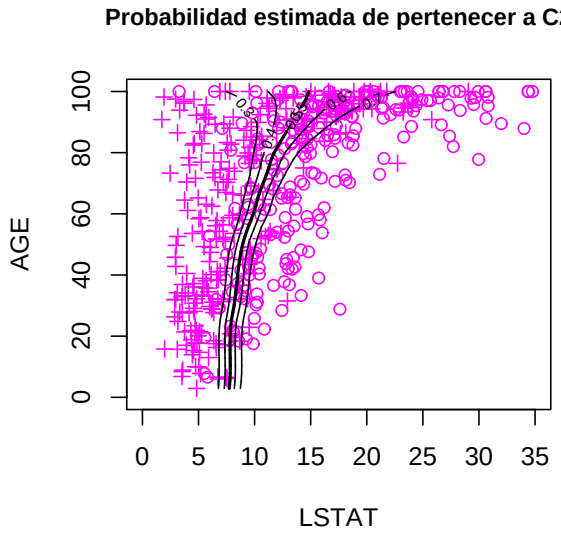


Figura 4.9: Regla discriminante basada en regresión logística local bivariante.

Otra forma de escribir esto es decir que

$$(Y_i|X = x_i) \sim N(m(x_i), \sigma^2),\ i = 1, \ldots, n.$$

El logaritmo de la verosimilitud de una función $\tilde{m}$, candidata a ser estimador de la función m desconocida es

$$l(\tilde{m}) = \log L(\tilde{m}) = C - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \tilde{m}(x_i))^2.$$

Con errores normales, maximizar la verosimilitud equivale a minimizar la suma de cuadrados de los residuos ajustados.

Para evitar elegir como estimador de m una función $\tilde{m}$ que interpole los datos $(\tilde{m}(x_i) = y_i,\ i = 1, \ldots, n)$, se maximiza localmente la verosimilitud de un modelo paramétrico:

$$l_{[t,w]}(\beta_0, \beta_1) = C - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1(x_i - t)))^2 w(x_i, t).$$

Los parámetros óptimos (alrededor de $x = t$ y usando los pesos dados por la función w) son $(\hat{\beta}_0(t), \hat{\beta}_1(t))$, y la función estimada es

$$\hat{m}(t) = \hat{\beta}_0(t).$$

Éste es otro modo de llegar al método de regresión lineal local.

Por su parte, en el modelo de regresión binaria local que hemos expuesto más arriba se tiene el modelo

$$(Y_i|X = x_i) \sim \text{Bernoulli}(m(x_{i1}, \ldots, x_{ip})).$$

Allí se estimó la función m mediante ajustes locales de modelos logísticos (con $\text{logit}(p_i)$ lineal en $(x_{i1}, \ldots, x_{ip})$) mediante máxima verosimilitud local.

En un problema de regresión no paramétrica general se quiere estimar la esperanza condicionada de una variable Y dado el valor x de una variable explicativa X (posiblemente multivariante)

$$m(x) = E(Y|X = x).$$

La propuesta basada en verosimilitud local consiste en proponer un modelo paramétrico que se ajuste bien localmente a la distribución de de $(Y|X = x)$:

$$(Y|X = x) \sim f(y; \theta(x)). \tag{4.7}$$

En este modelo habrá alguna forma estándar de calcular $E(Y|X = x)$ en función de los parámetros $\theta(x)$: $m(x) = g(\theta(x))$. Si se desea que $m(x)$ sea una función suave de x, implícitamente se está pidiendo que $\theta(x)$ también lo sea.

El modelo paramétrico local $f(y; \theta(x))$ debe tener en cuenta las características de $(Y|X = x)$. Por ejemplo, la variable Y puede ser una variable de conteo, o su soporte puede estar restringido a $\mathbb{R}^+$ o a $[0, 1]$. En esos caso, podrían usarse como modelos paramétricos locales los de Poisson, Gamma o Beta, respectivamente. En general, los modelos lineales generalizados son una buena colección entre la que buscar candidatos a modelos paramétricos locales.

La forma de ajustar el modelo paramétrico local es mediante máxima verosimilitud local, especificando una forma funcional sencilla para $\theta(x)$ como función de x. Concretamente, la modelización de $\theta(x)$ como un polinomio de primer grado en x da buenos resultados (es la que se ha usado en los casos de distribución condicionada normal o Bernoulli), aunque podría usarse un grado mayor que 1.

El logaritmo de la función de verosimilitud local alrededor de $x = t$ es

$$l_{[t,w]}(\beta_0, \beta) = \sum_{i=1}^{n} l_i(\beta_0, \beta) w(x_i, t),$$

donde $l_i(\beta_0, \beta) = \log f(y_i; \beta_0 + \beta^{\mathrm{T}}(x_i - t))$, $w(x_i, t)$ es el peso de la observación x_i en la stimación de $\theta(t)$, y se realiza un ajuste lineal local (se toma $\theta(x) =$

$\beta_0 + \beta^{\mathrm{T}}(x - t)$). El peso $w(x_i, t)$ vendrá dado usualmente por una función núcleo:

$$w(x_i, t) \propto K\left(\frac{x_i - t}{h}\right).$$

Los parámetros (β_0, β) así estimados serán $(\hat{\beta}_0(t), \hat{\beta}(t))$. Entonces,

$$\hat{\theta}(t) = \hat{\beta}_0(t) + \hat{\beta}(t)^{\mathrm{T}}(t - t) = \hat{\beta}_0(t),$$

y

$$\hat{m}(x) = \hat{E}(Y|X = x) = g(\hat{\theta}(t)) = g(\hat{\beta}_0(t)).$$

El modelo paramétrico ajustado localmente también proporciona estimaciones de la varianza del estimador local del parámetro θ,

$$V(\hat{\theta}(x)) = h(\theta(x)),\ \hat{V}(\hat{\theta}(x)) = h(\hat{\theta}(x)),$$

que pueden ser útiles para posteriores fases de la inferencia.

Por otra parte, además de la esperanza condicionada otras características de la distribución condicionada $(Y|X = x)$ también pueden ser calculadas a partir de $\theta(x)$. Por ejemplo, puede ser de interés estudiar la $V(Y|X = x)$ o un determinado quantil de $(Y|X = x)$.

La estimación no paramétrica mediante verosimilitud local que acabamos de exponer también puede sufrir los efectos de la maldición de la dimensionalidad si hay muchas variables explicativas en x (dimensión alta). Los *modelos aditivos generalizados* (ver Capítulo 6) consiguen evitar el problema de forma análoga a como los modelos aditivos lo eluden en el modelo de regresión múltiple no paramétrico: se pierde flexibilidad del modelo para ganar en capacidad de estimación y de interpretación. Los modelos aditivos generalizados extienden los modelos aditivos al caso en el que la variable respuesta no es continua o, en caso de serlo, ésta no sigue una distribución normal (dado el valor de la variable explicativa). Es el mismo tipo de extensión que permite pasar del modelo de regresión lineal múltiple al modelo lineal generalizado.

4.5. Inferencia en el modelo de regresión no paramétrica

En esta sección se listan algunos problemas de inferencia estadística que se pueden abordar usando estimación no paramétrica de la función de regresión. Los contenidos aquí expuestos se inspiran principalmente en los Capítulos 4, 5 y 6 del libro de Bowman y Azzalini (1997).

4.5.1. Bandas de variabilidad

En el modelo de regresión no paramétrico homocedástico (4.1), hemos visto que el estimador lineal local tiene sesgo

$$E(\hat{m}(x)) - m(x) = \frac{h^2 m''(x)\mu_2(K)}{2} + o(h^2)$$

y varianza

$$V(\hat{m}(x)) = \frac{R(K)\sigma^2(x)}{nhf(x)} + o\left(\frac{1}{nh}\right).$$

Esta varianza se puede estimar mediante

$$\hat{V}(\hat{m}(x)) = \frac{R(K)\hat{\sigma}^2(x)}{nhf(x)} + o\left(\frac{1}{nh}\right),$$

donde $\hat{\sigma}^2$ es alguno de los estimadores de σ^2 que se presentaron en la Sección 4.3.4, página 140.

Para un valor de h fijo, se puede probar que

$$\frac{\hat{m}(x) - E(\hat{m}(x))}{\sqrt{V(\hat{m}(x))}} \longrightarrow N(0,1) \text{ en ley cuando } n \longrightarrow \infty.$$

Esto permite construir intervalos de confianza para $E(\hat{m}(x))$, a los que llamaremos BANDAS DE VARIABILIDAD de $\hat{m}(x)$. Obsérvese que no son intervalos de confianza para $m(x)$. Si $\alpha = 0{,}05$,

$$IC_{1-\alpha} \equiv \left(\hat{m}(x) \mp 1{,}96\sqrt{\hat{V}(\hat{m}(x))}\right).$$

En los modelos de verosimilitud local, cada estimación local del modelo paramétrico da lugar a una estimación del parámetro local, $\hat{\theta}(x)$, y de su varianza, $\hat{V}(\hat{\theta}(x))$.

Teniendo en cuenta que

$$\hat{m}(x) = \hat{E}(Y|X = x) = g(\hat{\theta}(x))$$

y usando el método delta, se tiene el siguiente estimador de $V(\hat{m}(x))$:

$$\hat{V}(\hat{m}(x)) = \nabla g(\hat{\theta}(x))^{\mathrm{T}} \hat{V}(\hat{\theta}(x)) \nabla g(\hat{\theta}(x)),$$

donde $\nabla g(\theta)$ es el gradiente de la función g evaluado en el punto θ.

Obsérvese que las bandas de variabilidad son bandas puntuales para $E(\hat{m}(x))$. Las afirmaciones sobre la probabilidad (aproximada) de cobertura

del verdadero valor de $E(\hat{m}(x))$ son válidas cuando se realizan para un punto x concreto, pero no lo son cuando se hacen para todo x simultáneamente. Este tipo de afirmaciones serían unifores en x, y tendrían la forma

$$P\left(L(x) \leq E(\hat{m}(x)) \leq U(x), \text{ para todo } x \in \mathbb{R}\right) \approx 1-\alpha,$$

donde $L(x)$ y $U(x)$ serían funciones aleatorias (dependientes de la muestra observada).

En la Sección 5.7 de Wasserman (2006) se presenta una forma de construir bandas de confianza uniformes para $m(x)$.

4.5.2. Contraste de ausencia de efectos

En el modelo de regresión no paramétrico homocedástico (4.1),

$$Y_i = m(x_i) + \varepsilon_i, \ i = 1, \ldots, n,$$

nos planteamos contrastar la hipótesis nula de ausencia de efectos:

$$\begin{cases} H_0 : m(x) \text{ es constante e igual a } \mu_Y = E(Y), \\ H_1 : m(x) \text{ no es constante.} \end{cases}$$

Por analogía con el modelo de regresión lineal se propone usar como estadístico del contraste

$$F = \frac{(\mathrm{SCR}_0 - \mathrm{SCR}_1)/(\mathrm{gl}_0 - \mathrm{gl}_1)}{\mathrm{SCR}_1/\mathrm{gl}_1},$$

donde las sumas de cuadrados de los residuos (SCR) y los correspondientes grados de libertad (gl) son

$$\mathrm{SCR}_0 = \sum_{i=1}^{n} (y_i - \bar{y})^2, \ \mathrm{gl}_0 = n-1,$$

$$\mathrm{SCR}_1 = \sum_{i=1}^{n} (y_i - \hat{m}(x_i))^2,$$

con $\hat{m}(x)$ un estimador no paramétrico de $m(x)$, que tiene gl_1 grados de libertad efectivos, calculados como se vio en la ecuación (4.6).

Obsérvese que la distribución de F no tiene por qué ser una F de Snedecor, ni siquiera en el caso en el que los residuos del modelo tengan distribución normal.

La tabulación la distribución de F bajo la hipótesis nula se hace como en un test de permutaciones. Bajo H_0, cualquier permutación de $y_1, \ldots, y_n$ es igualmente probable para $x_1, \ldots, x_n$ fijo. Así, se realizan los siguientes pasos:

1. Se permuta $y_1, \ldots, y_n$ aleatoriamente y se obtiene $y_{i_1}, \ldots, y_{i_n}$. Se construye la muestra permutada

$$(x_j, y_{i_j}),\ j = 1, \ldots, n.$$

2. Se calcula el valor del estadístico F en la muestra permutada: F_P.

3. Se repiten B veces los pasos 1 y 2: $F_P^1, \ldots, F_P^B$.

4. Se compara el valor de F observado en la muestra original, F_{obs}, con $F_P^1, \ldots, F_P^B$, y se obtiene el p-valor del test:

$$p\text{-valor} = \frac{\#\{F_P^b > F_{obs}\}}{B}.$$

En el paso 2 se estima la función de regresión para cada muestra permutada. Si se representan las B funciones estimadas simultáneamente se obtiene una BANDA DE REFERENCIA DEL MODELO SIN EFECTOS, que permite contrastar gráficamente de forma aproximada la hipótesis nula de ausencia de efectos: si la función estimada en la muestra original sale fuera de la banda de referencia, se rechaza la hipótesis nula.

Hay otra forma de construir una BANDA DE REFERENCIA DEL MODELO SIN EFECTOS que no precisa de la obtención de muestras permutadas. Obsérvese que bajo la hipótesis nula ($m(x) = \mu_Y$, constante en x) el estimador local lineal es insesgado:

$$\hat{m}(x) = \sum_{i=1}^{n} w^*(x_i, x) y_i \Longrightarrow E(\hat{m}(x)) = \sum_{i=1}^{n} w^*(x_i, x)\mu_Y = \mu_Y = m(x).$$

Sea $\bar{y}$ la media muestral de $y_1, \ldots, y_n$. Éste es también un estimador insesgado de μ_Y. Así, para todo x,

$$E(\hat{m}(x) - \bar{y}) = 0,$$

$$V(\hat{m}(x) - \bar{y}) = V(\sum_{i=1}^{n} w^*(x_i, x) y_i - \sum_{i=1}^{n} (1/n) y_i) = \sigma^2 \sum_{i=1}^{n} (w^*(x_i, x) - (1/n))^2.$$

Teniendo en cuenta la normalidad asintótica, se tiene que

$$\left(\bar{y} \pm 1{,}96 \sqrt{\hat{\sigma}^2 \sum_{i=1}^{n} (w^*(x_i, x) - (1/n))^2} \right)$$

es una banda de referencia aproximada, de confianza 0,95, para la hipótesis nula de ausencia de efectos. Si la estimación no paramétrica $\hat{m}(x)$ sale fuera de esa banda H_0 debería rechazarse.

Debe recordarse siempre que un contraste gráfico tiene utilidad principalmente como herramienta descriptiva, y que es mucho menos preciso que un contraste basado en un test de permutaciones.

En el procedimeiento de contraste anterior se deja fijo el valor del parámetro de suavizado h en todas las estimaciones no paramétricas de $m(x)$ que se hacen usando las muestras permutadas que se generan en el paso 2. Por lo tanto, el p-valor del test y el resultado del contraste pueden depender del h elegido.

Por lo tanto es útil realizar gráficos de $(h, p\text{-valor}(h))$ que muestren si el resultado del contraste se mantiene inalterado para un gran rango de valores del parámetro de suavizado h, o por el contrario es grande la influencia de la elección de este parámetro.

4.5.3. Contraste de un modelo lineal

En el modelo de regresión no paramétrico homocedástico (4.1),

$$Y_i = m(x_i) + \varepsilon_i,\ i = 1, \ldots, n,$$

nos planteamos contrastar la hipótesis nula de que la función de regresión es lineal:

$$\begin{cases} H_0 : m(x) = \beta^{\mathrm{T}} x, \\ H_1 : m(x) \text{ no es lineal.} \end{cases}$$

Sea Y el vector de los n datos y_i, X la matriz de diseño y $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ la matriz sombrero. Así, los vectores de valores ajustados y de los residuos en el modelo lineal son, respectivamente,

$$\hat{Y}_L = HY,\ \hat{\varepsilon}_L = Y - \hat{Y}_L = (I_n - H)Y.$$

El contraste del modelo lineal se puede expresar como un contraste de ausencia de efectos en la relación entre los residuos estimados en el modelo lineal, $\hat{\varepsilon}_{L,i}$, y la variable explicativa x_i:

$$\begin{cases} H_0 : E(\hat{\varepsilon}_{L,i}) = 0, \\ H_1 : E(\hat{\varepsilon}_{L,i}) = m(x_i) - \beta^{\mathrm{T}} x_i. \end{cases}$$

Se aplica entonces todo lo expuesto en la Sección 4.5.2.

En particular, una banda de referencia aproximada, de confianza 0,95, para la hipótesis nula de linealidad viene dada por

$$\left(\hat{\beta}^{\mathrm{T}} x \pm 1{,}96 \sqrt{\hat{\sigma}^2 \sum_{i=1}^{n} (w^*(x_i, x) - h(x_i, x))^2} \right),$$

donde $h(x_i, x)$ es el elemento i-ésimo del vector fila

$h(x) = x(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, que hace $\hat{y}_x = \hat{\beta}^{\mathrm{T}}x = x^{\mathrm{T}}\hat{\beta} = x(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y = h(x)Y$.

Si la estimación no paramétrica $\hat{m}(x)$ sale fuera de esa banda H_0 debería rechazarse.

4.5.4. Contraste de un modelo lineal generalizado

Consideramos el modelo de regresión no paramétrica general (4.7) en el que se establece que la distribución condicionada de Y dado x sigue un modelo paramétrico local

$$(Y|X = x) \sim f(y; \theta(x))$$

con $\theta(x) \in \mathbb{R}$. Vamos a suponer que este modelo admite una parametrización en términos de $m(x) = E(Y|X = x) = g(\theta(x))$ y, quizás, de un parámetro de dispersión ψ, que no depende de x:

$$(Y|X = x) \sim f(y; m(x), \psi).$$

En el modelo de regresión con errores normales, ψ es la varianza residual σ^2. En el modelo de regresión logística este parámetro adicional ψ no aparece.

En este modelo nos planteamos contrastar la hipótesis nula de que el modelo adecuado es un modelo lineal generalizado, frente a que $m(x)$ sigue un modelo no paramétrico. Es decir, contrastar que la función $\theta(x)$ es lineal frente a que no lo es:

$$\begin{cases} H_0 : \theta(x) = \beta^{\mathrm{T}}x \ \ (\Longleftrightarrow m(x) = g(\beta^{\mathrm{T}}x)), \\ H_1 : \theta(x) \text{ no es lineal en } x. \end{cases}$$

Para realizar este contraste se propone usar un test de razón de pseudoverosimilitudes (*pseudo-likelihood ratio test*), que guarda analogía con el contraste de razón de verosimilitudes en los modelos paramétricos.

El estadístico del test será

$$\mathrm{PLRT} = 2\sum_{i=1}^{n}\left(\log f(y; \hat{m}(x_i), \hat{\psi}_{NP}) - \log f(y; g(\hat{\beta}^{\mathrm{T}}x_i), \hat{\psi}_{GLM})\right),$$

donde $\hat{m}(x)$ es un estimador no paramétrico de $m(x)$ (posiblemente el obtenido mediante verosimilitud local), $\hat{\psi}_{NP}$ es el estimador de ψ que proporciona la estimación no paramétrica de $m(x)$, mientras que $\hat{\beta}$ y $\hat{\psi}_{GLM}$ son las estimaciones de los parámetros ajustando el modelo lineal general que se propone en la hipótesis nula.

La tabulación de la distribución del estadístico PLRT bajo la hipótesis nula se hace mediante el procedimiento llamado BOOTSTRAP PARAMÉTRICO, que permite la generación de muestras aleatorias que verifican la hipótesis nula y guardan similitudes con la muestra observada. Se realizan los siguientes pasos:

1. Se estima el modelo lineal generalizado a partir de la muestra original: $\hat{\beta}$ y $\hat{\psi}_{GLM}$.
2. Se genera una muestra bootstrap: para cada valor de x_i, se simula y_i^* del modelo
$$(Y|X = x_i) \sim f(y; \hat{\beta}^{\mathrm{T}} x_i, \hat{\psi}_{GLM}).$$
3. Se calcula el valor del estadístico PLRT en la muestra bootstrap: PLRT^*.
4. Se repiten B veces los pasos 2 y 3: $\mathrm{PLRT}_1^*, \ldots, \mathrm{PLRT}_B^*$.
5. Se compara el valor de PLRT observado en la muestra original, PLRT_{obs}, con $\mathrm{PLRT}_1^*, \ldots, \mathrm{PLRT}_B^*$, y se obtiene el p-valor del test:
$$p\text{-valor} = \frac{\#\{\mathrm{PLRT}_b^* > \mathrm{PLRT}_{obs}\}}{B}.$$

Al igual que en el contraste de ausencia de efectos o en el de linealidad, también aquí es posible construir bandas alrededor de la estimación paramétrica que permiten hacer un contraste visual.

4.5.5. Igualdad de curvas de regresión

Supongamos que tenemos observaciones procedentes de I subpoblaciones, en cada una de las cuales los datos siguen el modelo de regresión no paramétrica (4.1):

$$y_{ij} = m_i(x_{ij}) + \varepsilon_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, I.$$

Nos planteamos contrastar la igualdad de las I curvas de regresión:

$$\begin{cases} H_0 : m_i(x) = m(x),\ i = 1, \ldots, I, \text{ para todo } x, \\ H_1 : \text{no todas las funciones de regresión son iguales.} \end{cases}$$

Se propone usar como estadístico del test el siguiente:

$$T_I = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (\hat{m}_i(x_{ij}) - \hat{m}(x_{ij}))^2}{\hat{\sigma}^2},$$

donde

- $\hat{m}(x)$ es la estimación no paramétrica de $m(x)$ bajo la hipótesis nula, es decir, usando todas las observaciones disponibles conjuntamente;
- $\hat{m}_i(x)$ es la estimación no paramétrica de $m(x)$ usando los datos de la subpoblación i, $i = 1, \ldots, I$;
- $\hat{\sigma}^2$ es la estimación de $\sigma^2 = V(\varepsilon_{ij})$, que se construye a partir de las estimaciones de σ^2 obtenidas al estimar $m(x)$ en cada subpoblación,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{I} \eta_i \hat{\sigma}_i^2}{\sum_{i=1}^{I} \eta_i},$$

siendo η_i el número de grados de libertad efectivos en la estimación de $m(x)$ en la subpoblación i-ésima, tal como se define en la ecuación (4.6).

Observar que el estadístico T_I responde a la estructura usual de los contrastes ANOVA: es el cociente de la *variabilidad entre subpoblaciones* dividida por la *variabilidad dentro de las subpoblaciones*.

Para tabular la distribución de T_I bajo la hipótesis nula hay varias alternativas. Una de ellas consiste en hacer un test de permutaciones. Si H_0 es cierta, pueden intercambiarse las etiquetas que indican la subpoblación a la que pertenece cada individuo sin por ello alterar la distribución del estadístico T_I. Así, se generan B muestras en las que estas etiquetas se permutan aleatoriamente y se calcula en cada una de ellas el estadístico T_I, obteniéndose los valores T_I^b, $b = 1, \ldots, B$. El p-valor del contraste será

$$p\text{-valor} = \frac{\#\{T_I^b > T_{I,obs}\}}{B}.$$

Otra forma de aproximar la distribución de T_I bajo H_0 es utilizar bootstrap:

1. Se calculan los residuos de los modelos no paramétricos estimados en cada subpoblación,

$$\hat{e}_{ij} = y_{ij} - \hat{m}_i(x_{ij}),\ j = 1, \ldots, n_i,\ i = 1, \ldots, I,$$

y se define el conjunto $E = \{\hat{e}_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, I\}$.

2. Se genera una muestra bootstrap así: para $j = 1, \ldots, n_i,\ i = 1, \ldots, I$,

$$y_{ij}^* = \hat{m}(x_{ij}) + \hat{e}_{ij}^*,$$

donde $\hat{e}_{ij}^*$ se extraen de E con reemplazamiento y de forma independiente.

3. Se calcula el valor del estadístico T_I en la muestra bootstrap: T_I^*.

4. Se repiten B veces los pasos 2 y 3: $T_{I,1}^*, \ldots, T_{I,B}^*$.

5. Se compara el valor de T_I observado en la muestra original, $T_{I,obs}$, con $T_{I,1}^*, \ldots, T_{I,B}^*$, y se obtiene el p-valor del test:

$$p\text{-valor} = \frac{\#\{T_{I,b}^* > T_{I,obs}^*\}}{B}.$$

En el caso de dos subpoblaciones ($I = 2$), el contraste anterior se puede complementar con un contraste gráfico aproximado. Se trata de construir una banda alrededor del estimador global $\hat{m}(x)$ de forma que si la hipótesis nula es cierta las estimaciones de $m(x)$ en ambas subpoblaciones caerán dentro de dicha banda con alta probabilidad.

Bajo la hipótesis nula, $d(x) = m_1(x) - m_2(x) = 0$ para todo x. Sea

$$\hat{d}(x) = \hat{m}_1(x) - \hat{m}_2(x)$$

el estimador de la función diferencia. Su varianza es

$$V(\hat{d}(x)) = V(\hat{m}_1(x)) + V(\hat{m}_2(x))$$

y puede ser estimada siguiendo las indicaciones dadas en la Sección 4.5.1.

Finalmente, para $\alpha = 0{,}05$, las bandas de aceptación de la hipótesis nula son

$$C(x) \equiv \left(\frac{1}{2}(\hat{m}_1(x) + \hat{m}_2(x)) \mp \frac{1{,}96}{2} \sqrt{\hat{V}(\hat{d}(x))} \right).$$

Es fácil comprobar que

$$\hat{m}_1(x) \notin C(x) \Longleftrightarrow \hat{m}_2(x) \notin C(x) \Longleftrightarrow |\hat{d}(x)| > 1{,}96 \sqrt{\hat{V}(\hat{d}(x))}.$$

Observar que estas bandas de aceptación son bandas puntuales (no son uniformes en x).

Las bandas anteriores sugieren utilizar un estadístico alternativo a T_I:

$$T_d = \int_{\mathbb{R}} \frac{(\hat{d}(x))^2}{\hat{V}(\hat{d}(x))} f(x) dx.$$

Su distribución bajo la hiótesis nula se puede aproximar mediante la generación de muestras permutadas o de muestras bootstrap.

Capítulo 5

Estimación de la regresión mediante splines

REFERENCIAS: Green y Silverman (1994). Fan y Gijbels (1996), Hastie, Tibshirani y Friedman (2001), Wasserman (2006).

5.1. Estimación mínimo cuadrática penalizada

Consideramos de nuevo el modelo de regresión no paramétrica (4.1):

$$y_i = m(x_i) + \varepsilon_i,\ i = 1, \ldots, n,$$

donde $\varepsilon_1, \ldots, \varepsilon_n$ son v.a. independientes con

$$E(\varepsilon_i) = 0, V(\varepsilon_i) = \sigma^2 \text{ para todo } i,$$

y los valores $x_1, \ldots, x_n$ son conocidos.

En el Capítulo 4 se propusieron unos estimadores de la función de regresión $m(x)$ (los ajustes por polinomios locales) y se estudiaron sus propiedades.

Ahora abordaremos la estimación de $m(x)$ de otra forma. Plantearemos un problema de optimización cuya solución dará lugar a una familia de estimadores no paramétricos.

Por analogía a la estimación mínimo cuadrática de un modelo paramétrico de regresión, podemos plantear la estimación de $m(x)$ como la resolución del problema de minimización de la suma de cuadrados de los residuos:

$$\min_{\tilde{m}:\mathbb{R}\longrightarrow\mathbb{R}} \sum_{i=1}^{n} (y_i - \tilde{m}(x_i))^2.$$

La solución a este problema es cualquier función que interpole los datos (x_i, y_i), $i = 1, \ldots, n$.

El problema de este planteamiento es que una función $\tilde{m}(x)$ que interpola los datos no es en general una función suave de x. Si queremos imponer que la solución $\tilde{m}(x)$ del problema de optimización tenga ciertas características de suavidad, hay que incluir en la función objetivo una penalización por falta de suavidad. Eso se consigue planteando el problema de mínimos cuadrados penalizados

$$\min_{\tilde{m}\in\mathcal{M}} \left\{ \sum_{i=1}^{n} (y_i - \tilde{m}(x_i))^2 + \phi(\tilde{m}) \right\}, \tag{5.1}$$

donde $\mathcal{M}$ es una clase de funciones suaves (por ejemplo, que tengan p derivadas continuas) y $\phi(\tilde{m})$ es un funcional ($\phi : \mathcal{M} \longrightarrow \mathbb{R}$) que penaliza la falta de suavidad de $\tilde{m}$.

Si los datos x_i están en un intervalo $[a, b] \subseteq \mathbb{R}$ una elección usual es tomar como $\mathcal{M}$ el espacio de las funciones de cuadrado integrable en $[a, b]$ con segunda derivada de cuadrado integrable en $[a, b]$,

$$\mathcal{M} = W_2^2[a,b] = \left\{ m : [a,b] \longrightarrow \mathbb{R} : \int_a^b (m(x))^2 dx < \infty, \text{ existe } m''(x) \text{ y } \right.$$

$$\left. \int_a^b (m''(x))^2 dx < \infty \right\},$$

y como funcional de penalización

$$\phi(m) = \lambda \int_a^b (m''(x))^2 dx, \ \lambda > 0.$$

El espacio $W_2^2[a, b]$ recibe el nombre de ESPACIO DE SOBOLEV DE SEGUNDO ORDEN EN $[a, b]$.

De este modo el problema (5.1) se escribe como

$$\min_{\tilde{m}\in W_2^2[a,b]} \left\{ \sum_{i=1}^{n} (y_i - \tilde{m}(x_i))^2 + \lambda \int_a^b (\tilde{m}''(x))^2 dx \right\}. \tag{5.2}$$

En la sección siguiente veremos que éste es un problema cuya solución es una FUNCIÓN SPLINE CÚBICA con NODOS en los valores observados de la variable explicativa $x_1, \ldots, x_n$.

5.2. Splines y splines cúbicos. Interpolación por splines

Definición 5.1 (Función spline) *La función $s : [a, b] \longrightarrow \mathbb{R}$ es una* FUNCIÓN SPLINE *(o un* SPLINE*) de grado p con nodos $t_1, \ldots, t_k$ si se verifica lo siguiente:*

1. *$a < t_1 < \cdots < t_k < b$ (denotaremos $t_0 = a$, $t_{k+1} = b$).*
2. *En cada intervalo $[t_j, t_{j+1}]$, $j = 0, \ldots, k$, $s(x)$ es un polinomio de grado p (o inferior).*
3. *La función $s(x)$ tiene $(p-1)$ derivadas continuas en $[a, b]$ (es decir, los polinomios que definen la función $s(x)$ en los intervalos $[t_{j-1}, t_j]$ y $[t_j, t_{j+1}]$ enlazan bien en t_j).*

Ejemplo 5.1

Splines cúbicos. Las funciones splines más comúnmente utilizadas son las de grado 3, o cúbicas. Son polinomios de grado 3 a trozos, que en los nodos son continuos con primera y segunda derivada continua. Se dice que el ojo humano no es capaz de apreciar discontinuidades en la derivada tercera (o superiores) de una función. Por lo tanto las funciones splines cúbicas representan adecuadamente el concepto poco formal de *función suave*.

.

Se dice que un spline es PERIÓDICO si $s(a) = s(b)$.

Se dice que un spline de grado p es NATURAL si p es impar, $p = 2l - 1$ con $l \geq 2$, y satisface que

$$s^{(l+j)}(a) = s^{(l+j)}(b) = 0,\ j = 0, 1, \ldots, l-1.$$

Obsérvese que éstas son $p + 1 = 2l$ restricciones.

Ejemplo 5.2

Splines cúbicos naturales. Si $p = 3$, entonces $l = 2$, y las 4 restricciones que debe verificar un spline cúbico para ser natural son éstas:

$$s''(a) = s''(b) = 0,\ s'''(a) = s'''(b) = 0.$$

Por lo tanto un spline cúbico natural $s(x)$ es lineal en $[a, t_1]$ y $[t_k, b]$. Además, $s''(t_1) = s''(t_k) = 0$.

Proposición 5.1 *Sea $S[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$ el conjunto de splines de grado p con nodos $t_1, \ldots, t_k$ definidos en $[a, b]$. $S[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$ es un espacio vectorial de dimensión $p + k + 1$.*

Demostración: El hecho de que $S[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$ sea un espacio vectorial se sigue de que es cerrado por sumas y por productos por escalares.

El cálculo de la dimensión de $S[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$ puede hacerse como sigue. Para definir uno de estos splines $s(x)$ hay que dar los $(p + 1)$ coeficientes que definen el polinomio de grado p en cada uno de los $(k + 1)$ intervalos en los que los nodos dividen el intervalo $[a, b]$. Por tanto intervienen $(p + 1)(k + 1) = pk + p + k + 1$ parámetros en la definición de $s(x)$.

Pero estos parámetros están sujetos a una serie de restricciones lineales que garantizan que los polinomios de grado p enlazan bien en cada nodo: las derivadas laterales de orden l, $l = 0, 1, \ldots, p - 1$, de $s(x)$ coinciden en t_j, $j = 1, \ldots, k$. Por tanto hay pk restricciones lineales.

Así, la dimensión de $S[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$ es

$$pk + p + k + 1 - pk = p + k + 1.$$

□

Ejemplo 5.3

Los splines cúbicos tiene dimensión $3 + k + 1 = k + 4$. Una base de $S[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$ viene dada así:

$$s_1(x) = 1, s_2(x) = x, s_3(x) = x^2, s_4(x) = x^3,$$

$$s_j(x) = (x - t_j)_+^3, \; j = 1, \ldots, k,$$

donde para cualquier número real u, $u_+ = \text{máx}\{0, u\}$ es la parte positiva de u.

Esta no es la única base de $S[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$, y de hecho hay otras bases que son más adecuadas para realizar cálculos numéricos (las bases de B-splines, por ejemplo, que veremos más adelante).

Proposición 5.2 *Sea $N[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$ el conjunto de splines naturales de grado p con nodos $t_1, \ldots, t_k$ definidos en $[a, b]$. $N[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$ es un espacio vectorial de dimensión k.*

Demostración: A las restricciones propias de $S[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$ hay que añadir las $2l = p + 1$ restricciones que deben verificar los splines naturales. Así, la dimensión de $N[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$ es

$$pk + p + k + 1 - pk - (p + 1) = k.$$

□

Proposición 5.3 *Dados (x_i, y_i), $i = 1, \ldots, n$, $n \geq 2$, $a < x_1 < \cdots < x_n < b$, existe un único spline natural $s(x)$ de grado p con nodos en x_i, $i = 1, \ldots, n$, que interpola esos datos:*

$$s(x_i) = y_i,\ i = 1, \ldots, n.$$

Demostración: Sea $\{s_1(x), \ldots, s_n(x)\}$ una base de $N[p; a, x_1, \ldots, x_n, b]$. Si $s(x) \in N[p; a, x_1, \ldots, x_n, b]$ entonces

$$s(x) = \sum_{j=1}^{n} \alpha_j s_j(x) \equiv (\alpha_1, \ldots, \alpha_n).$$

Cada restricción que $s(x)$ debe cumplir para interpolar los datos, $s(x_i) = y_i$, es una restricción lineal en los coeficientes α_j:

$$\sum_{j=1}^{n} \alpha_j s_j(x_i) = y_i,\ i = 1, \ldots, n.$$

Tenemos pues un sistema de n ecuaciones lineales con n incógnitas (α_j, $j = 1, \ldots, n$). La matriz de coeficientes de este sistema es

$$(s_j(x_i)),\ i = 1, \ldots, n,\ j = 1, \ldots, n,$$

que tiene rango completo (en caso contrario los elementos de la base no serían linealmente independientes). Ello implica que la solución del sistema existe y es única. Esa solución α_j^*, $j = 1, \ldots, n$, determina el único spline interpolador

$$s^*(x) = \sum_{j=1}^{n} \alpha_j^* s_j(x).$$

□

5.3. Suavizado por splines

Nos centraremos a partir de ahora en los splines cúbicos naturales. El siguiente resultado (cuya demostración ha sido extraída de Green y Silverman 1994) muestra que estas funciones tienen una propiedad de optimalidad que será útil más adelante.

Proposición 5.4 *Sea $n \geq 2$ y sea $s(x)$ el spline cúbico natural que interpola los datos (x_i, y_i), $i = 1, \ldots, n$, con $a < x_1 < \cdots < x_n < b$. Sea $g(x)$ otra función cualquiera de $\mathcal{M} = W_2^2[a, b]$ que también interpola los datos ($g(x_i) = y_i, i = 1, \ldots, n$). Entonces*

$$\int_a^b (s''(x))^2 dx \leq \int_a^b (g''(x))^2 dx$$

con igualdad si y sólo si $g(x) = s(x)$ para todo $x \in [a, b]$.

Demostración: Sea $h(x) = g(x) - s(x)$. Entonces $h(x_i) = 0$, $i = 1, \ldots, n$. Integrando por partes, se tiene que

$$I = \int_a^b s''(x)h''(x)dx = \left\{ \begin{array}{l} u = s''(x) \Longrightarrow du = s'''(x)dx \\ dv = h''(x)dx \Longrightarrow v = h'(x) \end{array} \right\} =$$

$$(h'(x)s''(x))|_a^b - \int_a^b h'(x)s'''(x)dx = -\int_a^b h'(x)s'''(x)dx.$$

La última igualdad se tiene porque $s''(a) = s''(b) = 0$, al ser $s(x)$ un spline natural. Por el mismo motivo $s'''(x)$ es nula en $[a, x_1)$ y $(x_n, b]$.

Por otra parte, por ser $s(x)$ spline cúbico $s'''(x)$ es constante entre cada par de nodos: $s'''(x) = s'''(x_i^+)$ si $x \in [x_i, x_{i+1})$, $i = 1, \ldots, n-1$. Así,

$$I = -\int_a^b h'(x)s'''(x)dx = -\sum_{i=1}^{n-1} s'''(x_i^+) \int_{x_i}^{x_{i+1}} h'(x)dx =$$

$$-\sum_{i=1}^{n-1} s'''(x_i^+)(h(x_{i+1}) - h(x_i)) = 0.$$

De lo anterior se deduce lo siguiente:

$$\int_a^b (g''(x))^2 dx = \int_a^b ((g''(x) - s''(x)) + s''(x))^2 dx =$$

$$\underbrace{\int_a^b (h''(x))^2 dx}_{\geq 0} + \int_a^b (s''(x))^2 dx + 2 \underbrace{\int_a^b s''(x)h''(x)dx}_{=I=0} \geq \int_a^b (s''(x))^2 dx.$$

La igualdad se da si y sólo si $\int_a^b (h''(x))^2 dx = 0$, lo que equivale a pedir que $h''(x) = 0$ para todo $x \in [a, b]$. Esto es equivalente a pedir que $h(x)$ sea lineal en $[a, b]$, pero como $h(x_i) = 0$, $i = 1, \ldots, n$ y $n \geq 2$, se tiene que la igualdad es equivalente a pedir que $h(x) = 0$ para todo $x \in [a, b]$, es decir, que $g(x) = s(x)$ para todo $x \in [a, b]$. □

Proposición 5.5 *Sea $n \geq 3$ y sean los datos (x_i, y_i), $i = 1, \ldots, n$, con $a < x_1 < \cdots < x_n < b$. Dado un valor del parámetro $\lambda > 0$, la solución del problema (5.2),*

$$\min_{\tilde{m} \in W_2^2[a,b]} \Psi(\tilde{m}) = \left\{ \sum_{i=1}^n (y_i - \tilde{m}(x_i))^2 + \lambda \int_a^b (\tilde{m}''(x))^2 dx \right\},$$

es un spline cúbico natural con nodos en $x_1, \ldots, x_n$.

Demostración: Sea $g(x) \in W_2^2[a, b]$ que no es spline cúbico natural con nodos en las x_i observadas. Sea $s_g(x)$ el spline cúbico natural con nodos en las x_i que interpola los puntos $(x_i, g(x_i))$, $i = 1, \ldots, n$. Así, $s_g(x_i) = g(x_i)$, $i = 1, \ldots, n$ y por tanto

$$\sum_{i=1}^n (y_i - g(x_i))^2 = \sum_{i=1}^n (y_i - s_g(x_i))^2.$$

Por otra parte, por la Proposición 5.4 se tiene que

$$\int_a^b (s_g''(x))^2 dx < \int_a^b (g''(x))^2 dx.$$

Así, $\Psi(s_g) < \Psi(g)$ y, por tanto, el óptimo de $\Psi(m)$ hay que buscarlo entre los splines cúbicos naturales con nodos en $x_1, \ldots, x_n$. □

Resumamos lo visto hasta aquí. La Proposición 5.5 garantiza que para buscar la función $\tilde{m}$ que soluciona el problema de optimización (5.2) basta con buscar entre las funciones que pertenecen a $N(3; a, x_1, \ldots, x_n, b)$, que sabemos que es un espacio vectorial de dimensión n. Si se fija una base $\{N_1(x), \ldots, N_n(x)\}$, de ese espacio vectorial, solucionar el problema (5.2) equivale a buscar las coordenadas $(\alpha_1, \ldots, \alpha_n)$ del elemento de $N(3; a, x_1, \ldots, x_n, b)$

que hace mínima la función objetivo de (5.2). Por lo tanto, hemos conseguido transformar un problema de optimización en un espacio de dimensión infinita, $W_2^2[a,b]$, en uno de optimización en dimensión n finita.

Vamos a ver que podemos encontrar la solución explícita del problema (5.2). Este problema, por la Proposición 5.5, se puede expresar así:

$$\min_{s\in N(3;a,x_1,\ldots,x_n,b)} \left\{ \sum_{i=1}^{n} (y_i - s(x_i))^2 + \lambda \int_a^b (s''(x))^2 dx \right\}. \tag{5.3}$$

Sea

$$s(x) = \sum_{j=1}^{n} \alpha_j N_j(x) = \alpha^{\mathrm{T}} \mathbf{N}(x),$$

donde $\alpha = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}}$ y $\mathbf{N}(x) = (N_1(x), \ldots, N_n(x))^{\mathrm{T}}$, la expresión de $s(x)$ en la base $\{N_1(x), \ldots, N_n(x)\}$ de $N(3; a, x_1, \ldots, x_n, b)$. Así,

$$s''(x) = \sum_{j=1}^{n} \alpha_j N_j''(x) = \alpha^{\mathrm{T}} \mathbf{N}''(x)$$

y

$$\int_a^b (s''(x))^2 dx = \int_a^b s''(x) s''(x)^{\mathrm{T}} dx =$$

$$\alpha^{\mathrm{T}} \int_a^b \mathbf{N}''(x)(\mathbf{N}''(x))^{\mathrm{T}} dx \alpha = \alpha^{\mathrm{T}} A \alpha,$$

donde A es una matriz $n \times n$ cuyo elemento (i,j) es

$$\int_a^b N_i''(x) N_j''(x) dx.$$

Sea $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$ y sea $\mathbf{N_x}$ la matriz $n \times n$ cuyo elemento (i,j) es $N_j(x_i)$. Entonces

$$\sum_{i=1}^{n} (y_i - s(x_i))^2 = (Y - \mathbf{N_x}\alpha)^{\mathrm{T}} (Y - \mathbf{N_x}\alpha).$$

Por lo tanto, el problema 5.3 puede reexpresarse como

$$\min_{\alpha\in\mathbb{R}^n} \Psi(\alpha) = (Y - \mathbf{N_x}\alpha)^{\mathrm{T}} (Y - \mathbf{N_x}\alpha) + \lambda \alpha^{\mathrm{T}} A \alpha, \tag{5.4}$$

que es resoluble explícitamente. En efecto,

$$\nabla \Psi(\alpha) = -2\mathbf{N_x}^{\mathrm{T}} (Y - \mathbf{N_x}\alpha) + 2\lambda A \alpha.$$

Igualando a 0 ese gradiente y despejando α, tenemos que el valor óptimo de α es

$$\hat{\alpha} = \left(\mathbf{N}_{\mathbf{x}}^{\mathrm{T}}\mathbf{N}_{\mathbf{x}} + \lambda A\right)^{-1}\mathbf{N}_{\mathbf{x}}^{\mathrm{T}}Y. \tag{5.5}$$

Por tanto, el vector de valores y_i ajustados será

$$\hat{Y} = \mathbf{N}_{\mathbf{x}}\hat{\alpha} = \mathbf{N}_{\mathbf{x}}\left(\mathbf{N}_{\mathbf{x}}^{\mathrm{T}}\mathbf{N}_{\mathbf{x}} + \lambda A\right)^{-1}\mathbf{N}_{\mathbf{x}}^{\mathrm{T}}Y = SY.$$

Es decir, el estimador spline es lineal en Y. Como consecuencia, es válido aquí todo lo discutido en la Sección 4.3 sobre la elección del parámetro de suavizado (en este caso el parámetro λ) por validación cruzada, validación cruzada generalizada, sobre número efectivo de parámetros (traza de S) y número efectivo de grados de libertad para la estimación de la varianza residual.

5.4. Propiedades del estimador spline de $m(x)$

Sea $\hat{m}_\lambda(x)$ el estimador spline de $m(x)$ cuando se usa λ como parámetro de suavizado. Se puede probar que si $\lambda \longrightarrow 0$ y $n\lambda^{1/4} \longrightarrow \infty$ cuando $n \longrightarrow \infty$, entonces

$$\mathrm{Sesgo}(\hat{m}_\lambda(x)) = O(\lambda),\ \mathrm{Var}(\hat{m}_\lambda(x)) = O\left(\frac{1}{n\lambda^{1/4}}\right),$$

$$\lambda_{\mathrm{AMSE}} = O\left(n^{-4/9}\right),\ \mathrm{MSE}(\lambda_{\mathrm{AMSE}}) = O\left(n^{-8/9}\right).$$

Se puede establecer una relación entre el estimador spline y un estimador núcleo de $m(x)$ con ventana variable: si $x \in (a, b)$,

$$\hat{m}_\lambda(x) \approx \frac{1}{nf(x)h(x)}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h(x)}\right) y_i =$$

$$\frac{\frac{1}{nh(x)}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h(x)}\right) y_i}{f(x)},$$

donde

$$K(u) = \frac{1}{2}e^{|u|/\sqrt{2}}\sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right),$$

que es un núcleo de orden 4, y

$$h(x) = \lambda^{1/4} f(x)^{-1/4}.$$

5.5. B-splines

En esta sección presentamos una base del espacio vectorioal de funciones splines de grado p con nodos $t_1, \ldots, t_k$ definidos en $[a, b]$, $S[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$, que presenta ventajas computacionales respecto a las bases del tipo presentado en el Ejemplo 5.3 para los splines cúbicos.

Se trata de las bases de B-splines, que se definen recursivamente. Sea $M = (p+1)$. Además de los k nodos $t_1, \ldots, t_k$ se definen $2M$ nodos auxiliares:

$$\tau_1 \leq \cdots \leq \tau_M \leq t_0,\ t_{k+1} \leq \tau_{k+M+1} \leq \cdots \leq \tau_{k+2M}.$$

La elección de estos nodos es arbitraria y puede hacerse

$$\tau_1 = \cdots = \tau_M = t_0,\ t_{k+1} = \cdots = \tau_{k+M+1} = \tau_{k+2M}.$$

Se renombran los nodos originales

$$\tau_{M+j} = t_j,\ j = 1, \ldots, k.$$

Se define la base de B-splines de orden 1 así:

$$B_{j,1} = I_{[\tau_j, \tau_{j+1}]},\ j = 1, \ldots, k + 2M - 1.$$

Para $m = 2, \ldots, M$, se definen los B-splines de orden m como

$$B_{j,m} = \frac{x - \tau_j}{\tau_{j+m-1} - \tau_j} B_{j,m-1} + \frac{\tau_{j+m} - x}{\tau_{j+m} - \tau_{j+1}} B_{j+1,m-1},$$

para $j = 1, \ldots, k + 2M - m$, entendiendo que uno de esos cocientes es 0 si el denominador es 0.

Se tiene que si $m = M = 4$ entonces las funciones $\{B_{j,m},\ j = 1, \ldots, k+4\}$ forman una base del conjunto de splines cúbicos con nodos $t_1, \ldots, t_k$ definidos en $[a, b]$, llamada BASE DE B-SPLINES CÚBICOS.

Ejemplo 5.4

La Figura 5.1 representa las 13 funciones que forman la base de B-splines cúbicos definidos en $[0, 1]$ con nueve nodos equiespaciados en 0,1, . . . , 0,9. Para construirla se tomaron

$$\tau_1 = \cdots = \tau_4 = 0,\ \tau_{14} = \cdots = \tau_{17} = 1.$$

.

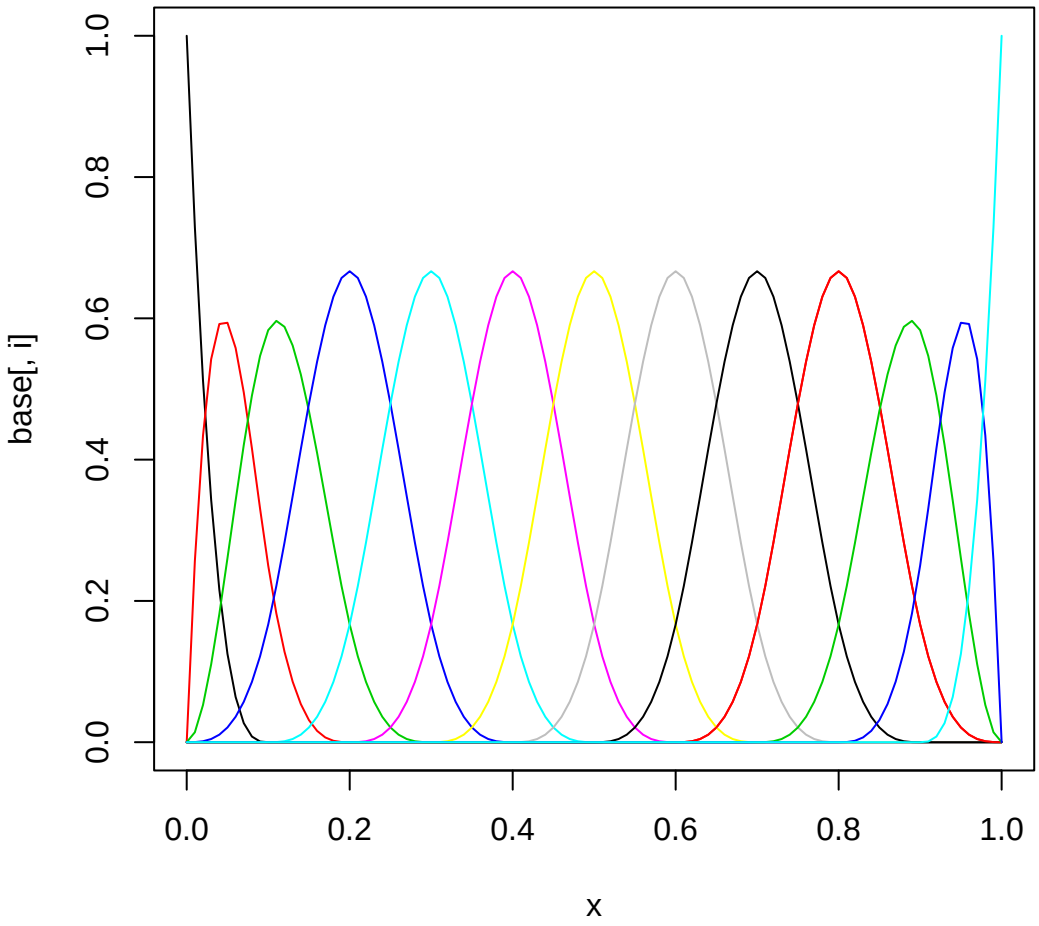


Figura 5.1: Base de B-splines cúbicos definidos en $[0, 1]$ con nueve nodos equiespaciados en 0,1, . . . , 0,9.

Entre las propiedades de los B-splines cúbicos, cabe destacar las siguientes:

1. $B_{j,4}(x) \geq 0$ para todo $x \in [a, b]$.
2. $B_{j,4}(x) = 0$ si $x \notin [\tau_j, \tau_{j+4}]$.
3. Si $j \in \{4, \ldots, k+1\}$, $B_{j,4}^{(l)}(\tau_j) = 0$, $B_{j,4}^{(l)}(\tau_{j+4}) = 0$, para $l = 0, 1, 2$.

La segunda de estas propiedades (cada base es distinta de 0 sólo en un pequeño subintervalo de $[a, b]$) es la responsable de las ventajas computacionales que presenta la base de B-splines cúbicos. Consideremos el problema análogo a (5.4), pero en el que la optimización se realiza en el conjunto de splines cúbicos (sin limitarse a aquellos que son naturales),

$$\min_{\beta \in \mathbb{R}^{n+4}} \Psi(\beta) = (Y - \mathbf{B}_\mathbf{x}\beta)^{\mathrm{T}}(Y - \mathbf{B}_\mathbf{x}\beta) + \lambda \beta^{\mathrm{T}} B \beta, \tag{5.6}$$

donde B es la matriz $(n+4) \times (n+4)$ cuyo elemento (i, j) es

$$\int_a^b B_i''(x) B_j''(x) dx,$$

y $\mathbf{B}_\mathbf{x}$ la matriz $n \times (n+4)$ cuyo elemento (i, j) es $B_j(x_i)$. Razonando igual que se hizo para llegar a la ecuación (5.5), se tiene que ahora la solución es

$$\hat{\beta} = \left(\mathbf{B}_\mathbf{x}^{\mathrm{T}} \mathbf{B}_\mathbf{x} + \lambda B\right)^{-1} \mathbf{B}_\mathbf{x}^{\mathrm{T}} Y. \tag{5.7}$$

Obsérvese que en este caso las matrices $\mathbf{B}_\mathbf{x}^{\mathrm{T}}\mathbf{B}_\mathbf{x}$ y B son *matrices banda*, con elementos (i, j) iguales a 0 si $|i - j| > 4$. Esto facilita la inversión de la matriz necesaria para determinar $\hat{\beta}$ (puede usarse la descomposición de Cholesky).

A partir de una base de B-splines cúbicos de $S[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$, con $k + 4$ elementos, es posible construir una base de B-splines cúbicos naturales, con k elementos que será base de $N[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$. Se hace

$$N_j = B_{j+2},\ j = 3, \ldots, k - 2.$$

Por otra parte, B_3, B_4, B_{k+1}, B_{k+2} se modifican para que sean splines cúbicos naturales, dando lugar a N_1, N_2, N_{k-1}, N_k, respectivamente. Se eliminan B_1, B_2, B_{k+3} y B_{k+4}.

En la práctica resulta prácticamente equivalente estimar el mejor spline cúbico (resolver el problema (5.6)) que estimar el mejor spline cúbico natural (resolver el problema (5.4)). La función de penalización hace que el mejor spline cúbico tenga un valor bajo (o nulo) fuera del intervalo $[x_1, x_n]$, lo que lleva a que ambos problemas tengan soluciones muy próximas.

En términos prácticos tampoco es necesario que se busque el mejor spline cúbico (o el mejor spline cúbico natural) con nodos en todos los x_i observados. De hecho esto hace que el coste computacional sea muy elevado si n es grande. Basta con tomar un número k de nodos suficientemente grande ($k = O(\log(n))$, por ejemplo) y tomarlos de forma que el nodo t_j sea el quantil $(j/(k+1))$ de los datos $x_1, \ldots, x_n$.

5.6. Ajuste de un modelo no paramétrico general

Supongamos que la variable aleatoria Y tiene una distribución condicionada a $X = x$ dada por

$$(Y|X = x) \sim f(y|\theta(x)),$$

donde $\theta(x) \in \mathbb{R}$ es una función suave de x. Dada una muestra

$$(x_i, y_i),\ i = 1, \ldots, n$$

acorde con este modelo, se puede plantear el problema de maximizar la verosimilitud penalizando por falta de suavidad:

$$\max_{\theta \in W_2^2[a,b]} \left\{ \sum_{i=1}^{n} \log(f(y_i|\theta(x))) + \lambda \int_a^b (\theta''(x))^2 dx \right\}. \tag{5.8}$$

Razonando como en el modelo de regresión, basta con optimizar en el conjunto de splines naturales con nodos en $x_1, \ldots, x_n$. En este caso, sin embargo, no se tendrá una solución cerrada del óptimo y habrá que recurrir a métodos numéricos de optimización.

La regresión binaria no paramétrica es un caso particular de este modelo.

Capítulo 6

Regresión múltiple y modelo aditivo generalizado

REFERENCIAS:
Hastie, Tibshirani y Friedman (2001) (secciones 9.1 y 11.2),
Fan y Gijbels (1996) (capítulo 7),
Bowman y Azzalini (1997) (capítulo 8),
Wasserman (2006) (sección 5.12), Hastie y Tibshirani
(1990), Ruppert, Wand y Carroll (2003)

6.1. Regresión múltiple

Hasta ahora hemos visto modelos de regresión con una variable de respuesta y una sola variable de entrada (regresor). La extensión del modelo de regresión no paramétrica al caso en el que hay p regresores es directa:

$$y_i = m(x_{i1}, \ldots, x_{ip}) + \varepsilon_i, \tag{6.1}$$

con $E(\varepsilon_i) = 0$ y $V(\varepsilon_i) = \sigma^2$, para $i = 1, \ldots, n$. Aquí la función de regresión m indica cómo varía y en función de la variable explicativa $x = (x_1, \ldots, x_p)$ de dimensión p.

Para definir en este marco los estimadores de la función de regresión mediante polinomios locales, necesitamos, por una parte, definir los pesos w_i de cada observación y, por otra, especificar qué variables explicativas se incluyen en cada modelo de regresión local.

La definición de los pesos w_i ha de seguir la misma lógica que en el caso univariante: si se quiere estimar r en el punto $t = (t_1, \ldots, t_p)$, los datos $(y_i; x_{i1}, \ldots, x_{ip})$ que más peso deben tener son aquéllos con valores de las variables explicativas $x = (x_1, \ldots, x_p)$ más cercanos a $t = (t_1, \ldots, t_p)$. Hay

que tener en cuenta que ahora las distancias entre x y t se deben medir en un espacio de dimensión p, y que hay muchas formas razonables de definir estas distancias.

Una forma sencilla de asignar pesos w_i que da buenos resultados en la práctica es la siguiente:

$$w_i = w(t, x_i) \propto \prod_{j=1}^{p} K\left(\frac{x_{ij} - t_j}{h_j}\right),$$

donde K es un núcleo univariante, h_j es un parámetro de suavizado adecuado para la j-ésima variable explicativa y el símbolo $\propto$ indica proporcionalidad. Si se toman núcleos gaussianos, esto equivale a asignar pesos alrededor de t usando como núcleo p-dimensional la función de densidad de una normal multivariante con p coordenadas independientes, cada una de ellas con varianza h_j^2 .

La definición de la distancia entre x_i y t será más precisa si se tiene en cuenta cómo es la relación de las variables explicativas entre sí. Para ello, en vez de tomar núcleos multivariantes con coordenadas independientes (es lo que ocurre si tomamos el producto de núcleos univariantes) se toma como núcleo la función de densidad de una variable aleatoria cuya matriz de varianzas y covarianzas sea un múltiplo h de la matriz de varianzas y covarianzas muestral C de los datos $(x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$. Por ejemplo, si se toma un núcleo gaussiano multivariante con estas características se tiene que

$$w_i = w(t, x_i) \propto \frac{1}{h^p} \exp\left\{-\frac{1}{2h}(x_i - t)^{\mathrm{T}} C^{-1} (x_i - t)\right\}.$$

La definición de los modelos de regresión polinómica ponderada que se ajustan localmente sigue también la lógica de lo expuesto en el caso univariante. Si se desea ajustar polinomios p-variantes de grado q, se deben incluir todos los términos posibles de la forma

$$\beta_{s_1 \cdots s_p} \prod_{j=1}^{p} (x_{ij} - t_j)^{s_j},$$

cuyo grado,

$$\sum_{j=1}^{p} s_j,$$

sea menor o igual que q. La estimación de la función de regresión será el término independiente del polinomio ajustado alrededor del punto t:

$$\hat{m}(t) = \hat{m}(t_1, \ldots, t_p) = \hat{\beta}_{0 \cdots 0}.$$

Por ejemplo, si hay dos variables explicativas el polinomio de grado 2 ajustado será

$$\beta_{00}+\beta_{10}(x_{i1}-t_1)+\beta_{01}(x_{i2}-t_2)+\beta_{11}(x_{i1}-t_1)(x_{i2}-t_2)+\beta_{20}(x_{i1}-t_1)^2+\beta_{02}(x_{i2}-t_2)^2$$

y la estimación de $m(t)$ en $t = (t_1, t_2)$ será $\hat{\beta}_{00}$, el estimador del término independiente del polinomio.

Ejemplo 6.1

La Figura 6.1 muestra la regresión no paramétrica bivariante de la variable `ROOM` (número medio de habitaciones por domicilio) como función de las variables `LSTAT` (porcentaje de población con estatus social en la categoría inferior) y `AGE` (porcentaje de viviendas construidas antes de 1940 en cada barrio de Boston). Se ha usado un núcleo producto de dos núcleos gaussianos univariantes. Los valores de los parámetros de suavizado son $h_1 = 2{,}5$ en la dimensión `LSTAT`, y $h_2 = 10$ en la dimensión `AGE`. Puede observarse que, para cualquier valor fijo de la variable `LSTAT`, el número de habitaciones tiene un máximo en un valor alto de la variable `AGE` (aproximadamente en el 75 % de viviendas anteriores a 1940). Posiblemente lo que ocurre es que las viviendas anteriores a 1940 eran en promedio más grandes que las construidas con posterioridad, y eso hace que el tamaño medio de las viviendas sea mayor en barrios con un porcentaje considerable de casas anteriores a esa fecha. El máximo local que la función de regresión estimada tiene en niveles altos de la primera variable explicativa y valores intermedios de la segunda, indica que la diferencia entre tamaños medios de las viviendas según la antigüedad de las mismas es más acentuada en los barrios pobres (valores altos de `LSTAT`) que en los ricos.

Tal como hemos expuesto el problema de la regresión múltiple no paramétrica, parece que no tiene características diferenciadas de la regresión simple. Sin embargo, la regresión múltiple plantea un problema específico difícil de solventar. Es el fenómeno conocido como *la maldición de la dimensionalidad* (*curse of dimensionality*, en inglés), que consiste en que en dimensiones altas en los entornos de un punto t casi no hay datos observados (esos entornos están casi vacíos). Ya nos ocupamos de este tema en el capítulo 3, dedicado a la estimación de la densidad.

La única forma de superar la maldición de la dimensionalidad es disponer de muestras de datos de enorme tamaño (esto suele ocurrir en problemas de minería de datos). Si éste no es el caso, hay que ser consciente de que el

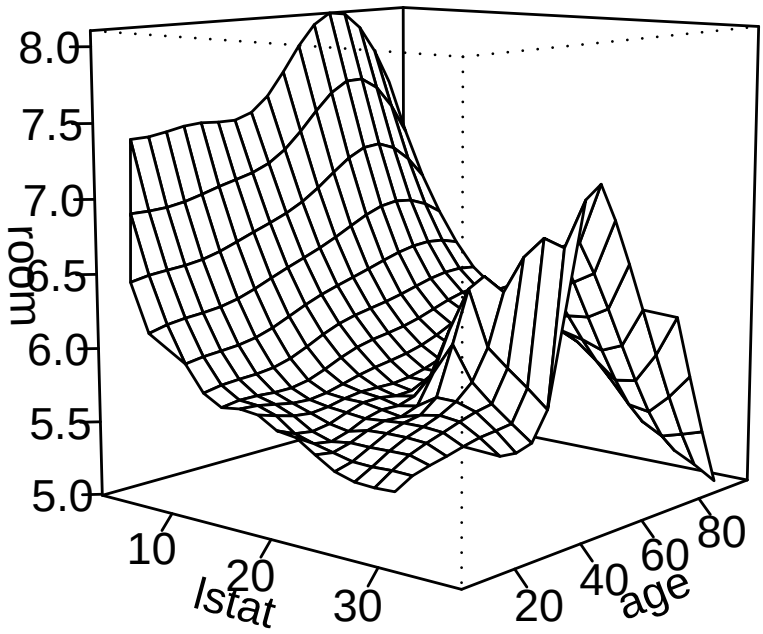


Figura 6.1: Regresión no paramétrica bivariante de la variable `ROOM` en función de las variables `LSTAT` y `AGE`.

comportamiento de los estimadores basados en polinomios locales se deteriora al aumentar el número de variables explicativas. Es recomendable no ir más allá de 3 o 4 dimensiones.

Existen métodos alternativos para estudiar la relación funcional entre la variable de respuesta y las variables explicativas a los que afecta menos la maldición de la dimensionalidad. Aquí expondremos únicamente los modelos aditivos y de la regresión mediante *projection pursuit*. En el Capítulo 7 de Fan y Gijbels (1996) o en la Sección 5.12 de Wasserman (2006) puede encontrarse información sobre otras posibilidades.

6.2. Modelos aditivos

Se plantea un modelo de regresión múltiple no paramétrico menos flexible que el que hemos visto hasta ahora. La pérdida en flexibilidad se ve compensada por el hecho de que el modelo es más fácilmente interpretable y se puede estimar con buenos resultados, incluso con alta dimensionalidad (muchas variables explicativas). El modelo aditivo es éste:

$$y_i = \alpha + \sum_{j=1}^{p} g_j(x_{ij}) + \varepsilon_i, \tag{6.2}$$

con $E(\varepsilon_i) = 0$ y $V(\varepsilon_i) = \sigma^2$ para todo $i = 1, \ldots, n$, y, además, $E(g_j(x_j)) = 0$ para todo $j = 1, \ldots, p$.

Las funciones $g_j(x_j)$ tendrán que ser estimadas no paramétricamente, puesto que no se especifica qué forma tienen. La única hipótesis adicional que se añade al modelo (6.1) es que las funciones $g_j(x_j)$ se combinan aditivamente para dar lugar a la función conjunta que relaciona la variable respuesta con las p variables explicativas. En cierto modo el modelo aditivo está a medio camino entre el modelo de regresión lineal múltiple paramétrico (que combina aditivamente transformaciones lineales de las variables, $\beta_j x_{ij}$) y el modelo de regresión múltiple no paramétrico.

Obsérvese que $E(y_i) = \alpha$ (ya que $E(\varepsilon_i) = 0$ y $E(g_j(x_j)) = 0$). Además, si el parámetro α y todas las funciones g_j fuesen conocidas, excepto la función g_k, entonces ésta podría estimarse mediante cualquier estimador no paramétrico univariante (por ejemplo, mediante un ajuste lineal local). Bastaría con aplicar ese estimador al conjunto de datos $(x_i, y_i^{(k)})$, donde

$$y_i^{(k)} = y_i - \alpha - \sum_{j=1, j\neq k}^{p} g_j(x_{ij}).$$

Esta observación lleva a proponer el algoritmo conocido como *backfitting* para estimar el modelo aditivo:

ALGORITMO Backfitting

Estimar α mediante $\hat{\alpha} = (1/n)\sum_{i=1}^{n} y_i$

Dar como estimaciones iniciales de las funciones g_k funciones cualesquiera $\hat{g}_k = g_k^0$, para $k = 1, \ldots, p$
(por ejemplo, $g_k^0(x_{ik}) = \hat{\beta}_k x_{ik}$, donde los coeficientes $\hat{\beta}_k$ son los estimados en el modelo de regresión lineal múltiple).

REPETIR

PARA CADA $k = 1, \ldots, p$,

estimar g_k mediante el ajuste no paramétrico univariante de los datos $(x_i, y_i^{(k)})$, donde

$$y_i^{(k)} = y_i - \hat{\alpha} - \sum_{j=1, j\neq k}^{p} \hat{g}_j(x_{ij}).$$

FIN PARA

HASTA convergencia.

FIN ALGORITMO

En Hastie y Tibshirani (1990) pueden encontrarse más detalles sobre este algoritmo y, en particular, sobre su convergencia y la unicidad de la solución a la que converge.

Ejemplo 6.2

Se ha ajustado un modelo aditivo a los datos de viviendas en los barrios de Boston. Se ha usado la librería `mgcv` del paquete R, que como estimador no paramétrico univariante usa el suavizado mediante splines. La Figura 6.2 muestra los resultados. Se ha representado la estimación de la función de regresión bivariante de `ROOM` sobre `LSTAT` y `AGE` en el panel superior. La comparación de este gráfico con la Figura 7 revela que el modelo aditivo no puede recoger el máximo local (situado alrededor de `LSTAT`=35, `AGE`=50) que vimos antes. Esto es una muestra de que este modelo es más rígido que el modelo de regresión no paramétrico.

En la Figura 6.3 se han representado las estimaciones no paramétricas de las contribuciones aditivas de cada variable explicativa, $g_{\texttt{LSTAT}}(\cdot)$ y $g_{\texttt{AGE}}(\cdot)$. En las etiquetas que acompañan a los ejes de ordenadas puede verse el número

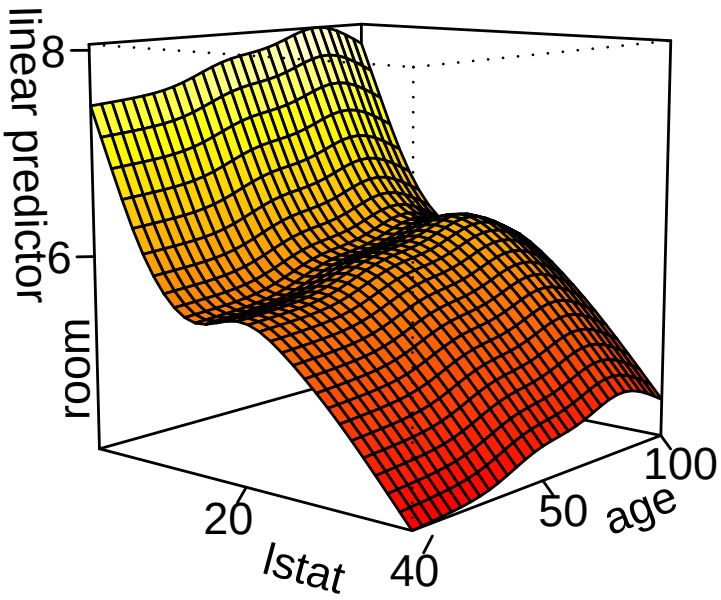


Figura 6.2: Modelo aditivo para el ajuste de `ROOM` como función de `LSTAT` y `AGE`. Función de regresión estimada.

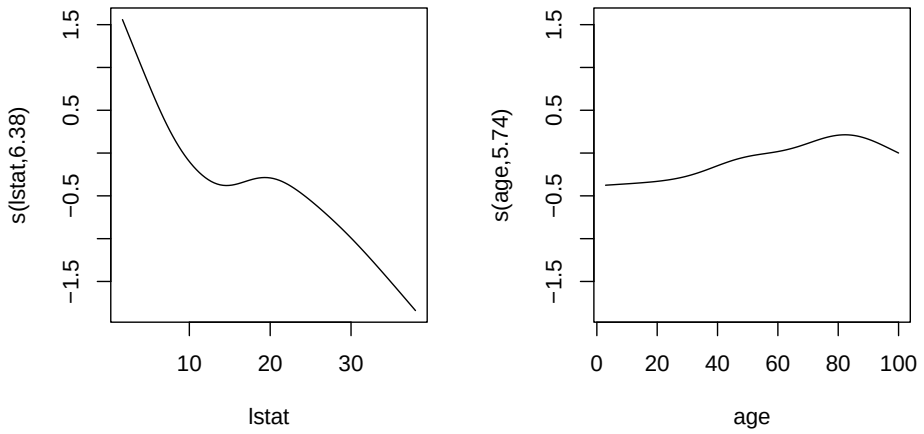


Figura 6.3: Estimaciones no paramétricas de las contribuciones aditivas $g_k(\cdot)$ de cada variable explicativa.

de parámetros equivalentes de ambas estimaciones. Se puede constatar que el gráfico de $g_{\texttt{LSTAT}}(\cdot)$ es (si se le suma la media global de `ROOM`) muy similar a la estimación de la función de regresión univariante de `ROOM`) sobre `LSTAT`.

Obsérvese que si se dan cortes al gráfico tridimensional, paralelos al plano (`LSTAT`, `ROOM`), los perfiles que se obtienen son copias de la función $g_{\texttt{LSTAT}}(\cdot)$. Análogo resultado se obtiene si los cortes son paralelos al plano (`AGE`, `ROOM`). De hecho la superficie representada se puede obtener desplazando la función $g_{\texttt{LSTAT}}(\cdot)$ en el espacio, apoyada sobre la función $g_{\texttt{AGE}}(\cdot)$ (o viceversa) y sumando la media global de `ROOM`.

.———————————————————————————————————————.

6.3. Regresión *projection pursuit*

El modelo de regresión basado en *projection pursuit* es el siguiente:

$$y_i = \alpha + \sum_{j=1}^{M} g_j(\alpha_j^T x_i) + \varepsilon_i, \tag{6.3}$$

con $E(\varepsilon_i) = 0$ y $V(\varepsilon_i) = \sigma^2$ para todo $i = 1, \ldots, n$, y cada α_j es un vector unitario (norma igual a 1) de $\mathbb{R}^p$. Además se supone que $E(g_j(\alpha^T x)) = 0$

para todo $j = 1, \ldots, M$.

Obsérvese que cada $z_j = \alpha_j^T x$ es la proyección del vector x en la dirección de α_j y que el modelo (6.3) puede reescribirse como un modelo aditivo en $z_1, \ldots, z_M$. Las direcciones α_j se buscan para que la variabilidad de y_i explicada por el modelo sea máxima. De ahí viene el nombre de *projection pursuit*, que podría traducirse como *búsqueda de la proyección*.

El siguiente algoritmo permite ajustar el modelo de regresión basado en *projection pursuit*:

Paso 1. Hacer $j = 0$, $\hat{\alpha} = \bar{y}_n$ e inicializar los residuos $\hat{\varepsilon}_i = y_i - \hat{\alpha}$.

Paso 2. Encontrar el vector unitario (la dirección) $\hat{\alpha}_j$ que minimiza

$$SCE(\alpha) = \sum_{i=1}^{n} (\hat{\varepsilon}_i - \hat{g}(\alpha_j^T x_i))^2,$$

donde $\hat{g}$ es un estimador no paramétrico de la regresión de $\hat{\varepsilon}_i$ sobre $\alpha_j^T x_i$. Llamar $\hat{g}_j$ a la función $\hat{g}$ correspondiente al valor óptimo $\hat{\alpha}_j$.

Paso 3. Actualizar los residuos $\hat{\varepsilon}_i = \hat{\varepsilon}_i - \hat{g}_j(\hat{\alpha}_j^T x_i)$ y hacer $j = j + 1$.

Paso 4. Volver al Paso 2 si no se cumplen las reglas de parada:

(a) Si $j = M$, parar.

(b) Si $SCE(\hat{\alpha}_j)/\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 < \delta$, parar.

Los valores M y/o δ se pueden elegir por validación cruzada.

6.4. Modelos aditivos generalizados

Supongamos que la variable aleatoria Y tiene una distribución condicionada a $X = (x_1, \ldots, x_p) \in \mathbb{R}^p$, dada por

$$(Y|X = (x_1, \ldots, x_p)) \sim f(y|\theta(x_1, \ldots, x_p)), \tag{6.4}$$

donde $\theta(x_1, \ldots, x_p) \in \mathbb{R}$ es una función suave de $x = (x_1, \ldots, x_p)$. Usualmente $\theta(x_1, \ldots, x_p)$ es una función biyectiva de $\mu(x_1, \ldots, x_p) = E(Y|X = (x_1, \ldots, x_p))$. Por ejemplo, si $(Y|X = (x_1, \ldots, x_p)) \sim \text{Bern}(\mu(x_1, \ldots, x_p))$ la función $\theta(x_1, \ldots, x_p)$ puede definirse a partir de μ mediante la transformación *logit*:

$$\theta(x_1, \ldots, x_p) = \log\left(\frac{\mu(x_1, \ldots, x_p)}{1 - \mu(x_1, \ldots, x_p)}\right).$$

De esta forma $\theta(x_1, \ldots, x_p)$ está libre de las restricciones que sí debe verificar $\mu(x_1, \ldots, x_p)$. Otras funciones *link* que suelen utilizarse para definir θ en función de μ son éstas: $\theta = \mu$ si $(Y|X = x)$ es Gaussiana (se tiene entonces el modelo aditivo), y $\theta = \log(\mu)$ si $(Y|X = x)$ es Poisson o Gamma.

La estimación del modelo 6.4 por verosimilitud local, tal y como se explicó en la Sección 4.4.2, se encuentra en este caso, en el que la variable explicativa es p-dimensional, con el problema de la *maldición de la dimensión* del mismo modo que éste aparece en la estimación del modelo de regresión múltiple no paramétrica (6.1) mediante la técnica de ajuste de polinomios locales. Para poder solventar este problema, se propone aquí una simplificación del modelo 6.4 análoga a la que da lugar a los modelos aditivos (6.2). Al modelo resultante le llamaremos **modelo aditivo generalizado** y se expresa como sigue:

$$(Y|X = (x_1, \ldots, x_p)) \sim f(y|\theta(x_1, \ldots, x_p)),\ \theta(x_1, \ldots, x_p) = \alpha + \sum_{j=1}^{p} g_j(x_j). \tag{6.5}$$

Obsérvese que si a este modelo se añade la restricción de que las funciones sean lineales se obtiene la formulación general del modelo lineal generalizado. Por lo tanto, el modelo aditivo generalizado está a medio camino entre el modelo de verosimilitud local multivariante 6.4 y el modelo lineal generalizado.

La forma de estimar un modelo aditivo generalizado combina el algoritmo backfitting (descrito en la Sección 6.2) con algún algoritmo de maximización de la verosimilitd usado en el ajuste de modelos lineales generalizados. En concreto, un algoritmo usado habitualmente en la estimación de modelos lineales generalizados es el basado en la iteración de ajustes por mínimos cuadrados reponderados. En este tipo de algoritmo, cada ajuste de una regresión múltiple por mínimos cuadrados ponderados se sustituye por el ajuste de un modelo aditivo ponderado mediante backfitting. De este modo se acaba ajustando el modelo aditivo generalizado en vez de ajustar el correspondiente modelo lineal generalizado. Vease el algoritmo contenido en la Figura 6.4.

El libro de Hastie y Tibshirani (1990) constituye la principal referencia sobre estos modelos, aunque se pueden encontrar introducciones a ellos en Hastie, Tibshirani y Friedman (2001) (sección 9.1) o Bowman y Azzalini (1997) (capítulo 8). La función `gam`, de la librería `mgcv` de R, permite ajustar modelos aditivos generalizados con gran flexibilidad.

Algorithm 9.2 *Local scoring algorithm for the additive logistic regression model.*

1. Compute starting values: $\hat{\alpha} = \log[\bar{y}/(1-\bar{y})]$, where $\bar{y} = \text{ave}(y_i)$, the sample proportion of ones, and set $\hat{f}_j \equiv 0\ \forall j$.

2. Define $\hat{\eta}_i = \hat{\alpha} + \sum_j \hat{f}_j(x_{ij})$ and $\hat{p}_i = 1/[1+\exp(-\hat{\eta}_i)]$.

   Iterate:

   (a) Construct the working target variable

   $$z_i = \hat{\eta}_i + \frac{(y_i - \hat{p}_i)}{\hat{p}_i(1-\hat{p}_i)}.$$

   (b) Construct weights $w_i = \hat{p}_i(1-\hat{p}_i)$

   (c) Fit an additive model to the targets z_i with weights w_i, using a weighted backfitting algorithm. This gives new estimates $\hat{\alpha}, \hat{f}_j, \forall j$

3. Continue step 2. until the change in the functions falls below a prespecified threshold.

Figura 6.4: Algoritmo para la estimaación de un modelo aditivo generalizado. (Fuente: Hastie, Tibshirani y Friedman, 2001)

6.5. Modelos semiparamétricos

En ocasiones algunas de las variables explicativas que intervienen en la definición de un modelo aditivo generalizado (6.5) (o en un modelo aditivo (6.2)) afectan a la variable respuesta de forma lineal. Si esto es conocido, el modelo (6.5) puede reformularse permitiendo que algunas de las funciones $g_j(x_j)$ sean lineales: $g_j(x_j) = \beta_j(x_j)$. Otras posibles modificaciones del modelo (6.5) son las siguientes:

- Estimar noparamétricamente el efecto conjunto de dos (o más) variables explicativas. Ello supone, por ejemplo, sustituir $g_j(x_j) + g_h(x_h)$ por $g_{j,h}(x_j, x_h)$ en (6.5).
- Estimar el efecto de una variable x_j de forma diferente en cada una de las clases determinadas por otra variable categórica x_h. Estos efectos podrías ser estimados lineal o noparamétricamente.

Los modelos obtenidos al incorporar estas modificaciones al modelo (6.5) se conocen como modelo semiparamétricos y pueden ser ajustados usando la función `gam` la librería `mgcv` de R.

El libro de Ruppert, Wand y Carroll (2003) está dedicado a los modelos semiparamétricos. Estos autores han desarrollado paralelamente la librería `SemiPar` de R, que permite ajustar estos modelos. Concretamente la función `spm` tiene similitudes con la función `gam`, aunque incorpora algunas opciones nuevas.

Apéndice A

Algunos conceptos y resultados matemáticos

Definición A.1 (Límites superior e inferior de una sucesión) *Sea* $\{x_n\}_n$, *una sucesión de números reales. Se define su límite superior como*

$$\lim_n \sup = \inf_n \sup_{m \geq n} x_n = \lim_n \sup_{m \geq n} x_n.$$

Se define su límite inferior como

$$\lim_n \inf = \sup_n \inf_{m \geq n} x_n = \lim_n \inf_{m \geq n} x_n.$$

Definición A.2 (Órdenes de convergencia de sucesiones) *Sean* $\{x_n\}_n$, $\{y_n\}_n$ *dos sucesiones de números reales. Cuando* n *tiende a* ∞

1. $x_n = O(y_n) \iff \lim\sup_n \left|\frac{x_n}{y_n}\right| < \infty$.
2. $x_n = o(y_n) \iff \lim_n \left|\frac{x_n}{y_n}\right| = 0$.

Definición A.3 (Órdenes de convergencia de sucesiones de v.a.) *Sean* $\{X_n\}_n$, $\{Y_n\}_n$ *dos sucesiones de variables aleatorias. Cuando* n *tiende a* ∞

1. $x_n = O_p(y_n)$ *si y sólo si para todo* $\varepsilon > 0$ *existen* $\delta > 0$ *y* $N \in \mathbb{N}$ *tales que para todo* $n \geq N$
$$P\left(\left|\frac{X_n}{Y_n}\right| > \delta\right) < \varepsilon,$$
es decir, si y sólo si $|X_n/Y_n|$ *está acotada en probabilidad.*

2. $X_n = o_p(Y_n)$ *si y sólo si para todo* $\varepsilon > 0$

$$\lim_n P\left(\left|\frac{x_n}{y_n}\right| > \varepsilon\right) = 0,$$

es decir, si y sólo si X_n/Y_n *tiende a 0 en probabilidad.*

Teorema A.1 (Teorema de Taylor) *Sea* $f(x)$ *una función con* $(r+1)$ *derivadas en un intervalo* $I \subseteq \mathbb{R}$*. Para cada par de puntos* x, a *de* I *se tiene que*

$$f(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \cdots$$
$$\cdots + \frac{1}{r!}f^{(r)}(a)(x-a)^r + \frac{1}{(r+1)!}f^{(r+1)}(\alpha(x,a))(x-a)^{r+1}$$

donde $\alpha(x,a)$ *es un punto intermedio entre* x *y* a *(luego* $|\alpha(x,a)-a| \leq |x-a|$*). Al último sumando se le llama* RESTO DE TAYLOR*, se le denota por* $R_{f,r}(x,a)$ *y admite otras expresiones:*

$$R_{f,r}(x,a) = \frac{1}{(r+1)!}f^{(r+1)}(\tilde{\alpha}(x,a))(x-a)(x-\tilde{\alpha}(x,a))^r,$$

y, si $f^{(r+1)}$ *es integrable en* I*,*

$$R_{f,r}(x,a) = \int_a^x \frac{1}{r!}f^{(r+1)}(t)(x-t)^r dt$$

(suponiendo que $a < x$*).*

Corolario A.1 *Si* $f^{(r+1)}$ *es acotada en* I *entonces*

$$R_{f,r}(x,a) = o((x-a)^r) \text{ cuando } (x-a) \longrightarrow 0,$$
$$R_{f,r}(x,a) = O((x-a)^{r+1}) \text{ cuando } (x-a) \longrightarrow 0.$$

Teorema A.2 (Teorema del Valor Medio Integral) *Si* f *es continua en* $[a,b]$*, entonces*

$$\int_a^b f(t)dt = f(\psi)(b-a)$$

para algún $\psi \in [a,b]$*.*

Teorema A.3 (Teorema del Valor Medio Integral Generalizado) *Si* f *y* g *son continuas en* $[a,b]$ *y* $g(t) \geq 0$ *para todo* $t \in [a,b]$*, entonces*

$$\int_a^b f(t)g(t)dt = f(\psi)\int_a^b g(t)dt$$

para algún $\psi \in [a,b]$*.*

Definición A.4 (Consistencia en Error Cuadrático Medio) *Un estimador $\hat{\theta}_n = \hat{\theta}(X_1 \ldots, X_n)$ es consistente en Error Cuadrático Medio para el parámetro θ si*

$$\lim_{n \longrightarrow \infty} MSE(\hat{\theta}_n) = 0,$$

donde MSE, el Error Cuadrático Medio (Mean Square Error, en inglés), se define como

$$MSE(\hat{\theta}_n) = E[(\hat{\theta}_n - \theta)^2] = (Sesgo(\hat{\theta}_n))^2 + V(\hat{\theta}_n),$$

donde $Sesgo(\hat{\theta}_n) = E(\hat{\theta}_n) - \theta$ es el sesgo de $\hat{\theta}_n$ como estimador de θ.

La desigualdad de Chebychev garantiza el siguiente resultado.

Proposición A.1 *Si $\hat{\theta}_n$ es un estimador consistente en Error Cuadrático Medio de θ entonces*

$$\hat{\theta}_n \longrightarrow_n \theta \text{ en probabilidad.}$$

Referencias

Bowman, A. W. y A. Azzalini (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford: Oxford University Press.

Cristóbal, J.A (1992). *Inferencia Estadística*. Universidad de Zaragoza.

Delicado, P. (2006). Local likelihood density estimation based on smooth truncation. *Biometrika*, **93**, 472–480. http://www-eio.upc.es/~delicado/research.html.

Fan, J. y I. Gijbels (1996). *Local polynomial modelling and its applications*. London: Chapman & Hall.

García-Nogales, A. (1998). *Estadística Matemática*. Universidad de Extremadura.

Gasser, T., L. Sroka y C. Jennen-Steinmetz (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* (3), 625–633.

Gibbons, J. D. (1993a). *Nonparametric Measures of Association*. Number 07-091 in Sage University Papers series on Quantitative Applications in the Social Sciences. Newbury Park, CA: Sage.

Gibbons, J. D. (1993b). *Nonparametric Statistics: An Introduction*. Number 07-090 in Sage University Papers series on Quantitative Applications in the Social Sciences. Newbury Park, CA: Sage.

Gibbons, J. D. (1997). *Nonparametric Methods for Quantitative Analysis* (Third ed.). Syracuse, N.Y.: American Sciences Press.

Gibbons, J. D. y S. Chakraborti (1992). *Nonparametric Statistical Inference* (Third ed.). New York: Marcewl Dekker.

Green, P.J. y B.W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman & Hall.

Hastie, T., R. Tibshirani y J. Friedman (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer.

Hastie, T. J. y R. J. Tibshirani (1990). *Generalized additive models*. Monographs on Statistics and Applied Probability. London: Chapman and Hall Ltd.

Hollander, M. y D. A. Wolfe (1999). *Nonparametric Statistical Methods* (Second ed.). Wiley & Sons.

Leach, C. (1982). *Fundamentos de Estadística. Enfoque no Paramétrico para Ciencias Sociales*. México: Limusa.

Pratt, J. W. y J. D. Gibbons (1981). *Concepts of Nonparametric Theory*. New York: Springer-Verlag.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* (4), 1215–1230.

Ruppert, D., S. J. Sheather y M. P. Wand (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* (432), 1257–1270.

Ruppert, D., M. P. Wand y R. J. Carroll (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley.

Scott, D. W., R. A. Tapia y J. R. Thompson (1977). Kernel density estimation revisited. *Nonlinear Anal.*, **1**, 339–372.

Sheather, S. J. y M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B, Methodological*, **53**, 683–690.

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

Simonoff, J. S. (1996). *Smoothing methods in statistics*. New York: Springer.

Vélez, R. y A. García (1993). *Principios de Inferencia Estadística*. UNED.

Wand, M. P. y M. C. Jones (1995). *Kernel smoothing*. London: Chapman and Hall.

Wasserman, L. (2006). *All of Nonparametric Statistics*. New York: Springer.