

En muchas situaciones interesa analizar la relación existente entre dos variables,  $X$  e  $Y$ . El análisis de regresión estudia de que forma  $Y$  (la variable dependiente) se puede explicar a partir de  $X$ . Si  $Y$  depende de  $X$  entonces

$$Y = m(X),$$

donde  $m$  es una función. En muchos casos no existe ninguna teoría que diga cómo debe de ser  $m$ . El análisis de la información empírica disponible nos debería de proporcionar información sobre  $m$ .

Consideremos el siguiente ejemplo. Sea  $Y$  el gasto en patatas y  $X$  los ingresos netos de una familia. Nos interesa saber cuánto es el gasto en patatas dado un nivel de ingresos. ¿Cómo es la función  $m$  que relaciona los ingresos  $X$  con el gasto  $Y$ ? En teoría económica se dice que un producto es *inferior* si el nivel de gasto tiende a disminuir a partir de un cierto nivel de ingresos.

¿Son las patatas un producto *inferior*? Para saberlo deberíamos recoger un conjunto representativo de datos y estimar  $m$ . Como las leyes generales pueden no ser válidas para un consumidor particular debemos tener en cuenta que la relación puede ser sólo válida en *términos medios*, es decir,

$$y_i = m(x_i) + \epsilon_i, \quad i = 1, \dots, n,$$

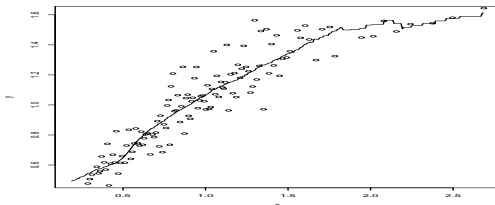
de forma que

$$m(x) = \mathbb{E}(Y|X = x).$$

La variable  $\epsilon_i$  representa la variabilidad en el gasto del individuo  $i$  sobre el consumo medio correspondiente a su nivel de ingresos  $x_i$ .

El objetivo es estimar la función  $m$  a partir de la observación de un conjunto finito de observaciones

$$(x_i, y_i), \quad i = 1, \dots, n.$$



En regresión paramétrica habitualmente se supone que  $m$  depende linealmente de un vector de parámetros. Por ejemplo, en regresión lineal simple se supone que

$$m(x) = \alpha + \beta x.$$

Este modelo sería muy restrictivo para ejemplos como el anterior. No permite que el consumo se incremente hasta un cierto nivel a partir del cual baja o se mantiene estable. En los modelos de regresión no paramétricos no se impone ninguna restricción a priori sobre  $m$ . Obviamente existe un precio a pagar por esta flexibilidad.

Antes de presentar los diferentes métodos de estimación es conveniente recordar brevemente el concepto de media condicional. Si  $X$  e  $Y$  son dos variables aleatorias con función de densidad conjunta  $f(x, y)$  se define la media condicional de  $Y$  dado  $X = x$  como

$$\mathbb{E}(Y|X = x) = \int y f(y|x) dy = \frac{\int y f(x, y) dy}{f_X(x)},$$

donde

$$f(y|x) = \frac{f(x, y)}{f_X(x)}$$

es la densidad condicional de  $Y$  dado  $X = x$  y  $f_X$  es la densidad marginal de  $X$ .

## Ejercicio

Supongamos que

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left( \begin{pmatrix} \mu \\ \eta \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix} \right),$$

con densidad

$$\frac{1}{2\pi\sigma\tau\sqrt{1-\rho^2}} \exp \left\{ \frac{-\left(\frac{x-\mu}{\sigma}\right)^2 - \left(\frac{y-\eta}{\tau}\right)^2 + 2\rho\left(\frac{x-\mu}{\sigma}\right)\left(\frac{y-\eta}{\tau}\right)}{2(1-\rho^2)} \right\}$$

Demuestra que  $\mathbb{E}(Y|X=x) = \alpha + \beta x$  donde  $\alpha = \eta - \mu\rho\tau/\sigma$  y  $\beta = \rho\tau/\sigma$ .

Nos centraremos en el caso de diseño aleatorio. En este diseño se supone que se dispone de una muestra aleatoria simple

$$(X_1, Y_1), \dots, (X_n, Y_n), \quad i = 1, \dots, n$$

de la distribución conjunta  $(X, Y)$  con densidad  $f(x, y)$ .

En algunos contextos (en ciencias principalmente) el investigador puede diseñar previamente el experimento y fijar de antemano los valores de la variable  $X$ . En este caso  $X$  no es una variable aleatoria mientras que  $Y$  sí. Esto simplifica la estimación de  $m$  así como el análisis de las propiedades estadísticas de los estimadores.



El estimador tipo núcleo se basa en la fórmula de la media condicional

$$\mathbb{E}(Y|X = x) = \frac{\int y f(x, y) dy}{f_X(x)}.$$

Para estimar  $m$  basta por tanto estimar  $f_X(x)$  y  $f(x, y)$ . La estimación tipo núcleo de  $f_X(x)$  y  $f(x, y)$  ya la hemos visto. Recordemos que para estimar la densidad bivalente  $f(x, y)$  es habitual emplear el estimador tipo núcleo con núcleo producto

$$\hat{f}_{n,K}(x, y) = \frac{1}{n} \sum_{j=1}^n K_h(x - X_j) K_g(y - Y_j).$$

Por tanto el estimador del numerador de la media condicional sería

$$\begin{aligned} \int y \hat{f}_{n,K}(x, y) dy &= \frac{1}{n} \sum_{j=1}^n \frac{1}{h} K\left(\frac{x - X_j}{h}\right) \int \frac{y}{g} K\left(\frac{y - Y_j}{g}\right) dy \\ &= \frac{1}{n} \sum_{j=1}^n \frac{1}{h} K\left(\frac{x - X_j}{h}\right) \int (zg + Y_j) K(z) dz = \frac{1}{n} \sum_{j=1}^n K_h(x - X_j) Y_j, \end{aligned}$$

donde en la última igualdad se ha usado que  $K$  es una función simétrica alrededor del cero con integral uno. El estimador de la función  $m$  resultante de reemplazar las cantidades desconocidas por sus estimadores en la fórmula de la esperanza condicional fue propuesto por Nadaraya y Watson en 1964

$$\hat{m}_{n,K}(x) = \frac{\sum_{j=1}^n K_h(x - X_j) Y_j}{\sum_{k=1}^n K_h(x - X_k)}$$

i) ¿Cómo es el estimador si utilizamos para estimar la densidad el estimador Naive? Recordemos que el estimador Naive es un estimador tipo núcleo con núcleo

$$\omega(x) = \frac{1}{2}\mathbb{I}(-1 \leq x \leq 1).$$

Utilizando este núcleo el estimador de Nadaraya-Watson en el punto  $x$  es la media de aquellos valores  $Y_j$  para los cuales su correspondiente  $X_j$  esté en el intervalo  $(x - h, x + h)$

$$\hat{m}_{n,\omega}(x) = \frac{\sum_{j=1}^n \mathbb{I}(X_j \in (x - h, x + h)) Y_j}{\sum_{k=1}^n \mathbb{I}(X_j \in (x - h, x + h))}$$

ii) El estimador de Nadaraya-Watson se puede reescribir como

$$\hat{m}_{n,K}(x) = \sum_{j=1}^n \frac{K_h(x - X_j)}{\sum_{k=1}^n K_h(x - X_k)} Y_j = \sum_{j=1}^n W_{hj}(x) Y_j,$$

donde

$$W_{hj}(x) \equiv W_j(x) = \frac{K_h(x - X_j)}{\sum_{k=1}^n K_h(x - X_k)} = \frac{K\left(\frac{x - X_j}{h}\right)}{\sum_{k=1}^n K\left(\frac{x - X_k}{h}\right)}.$$

Por tanto el estimador tipo núcleo de la función de regresión es una media (local) ponderada de los valores observados de la variable  $Y$  donde  $\sum_{j=1}^n W_j(x) = 1$ .

iii) El estimador tipo núcleo es un caso particular de los denominados estimadores lineales. Los estimadores lineales son aquellos estimadores  $\hat{m}$  que se pueden escribir como

$$\hat{m}(x) = \sum_{j=1}^n l_j(x) Y_j$$

donde  $\sum_{j=1}^n l_j(x) = 1$ . Por tanto el estimador tipo núcleo de la función de regresión es un estimador lineal con

$$l_j(x) = \frac{K\left(\frac{x-X_j}{h}\right)}{\sum_{k=1}^n K\left(\frac{x-X_k}{h}\right)}, \quad j = 1, \dots, n.$$

iv) Podemos utilizar el cálculo matricial para evaluar el estimador de Nadaraya-Watson en una rejilla de puntos  $(t_1, \dots, t_m)$ . El valor del estimador en el punto  $t_i$  viene dado por

$$\hat{m}_{n,K}(t_i) = \sum_{j=1}^n \frac{K_h(t_i - X_j)}{\sum_{k=1}^n K_h(t_i - X_k)} Y_j = \sum_{j=1}^n L_{ij} Y_j, \quad i = 1, \dots, m$$

donde

$$L_{ij} = W_j(t_i) = \frac{K\left(\frac{t_i - X_j}{h}\right)}{\sum_{k=1}^n K\left(\frac{t_i - X_k}{h}\right)}, \quad i = 1, \dots, m, j = 1, \dots, n$$

Por tanto, si  $\hat{\mathbf{m}} = (\hat{m}(t_1), \dots, \hat{m}(t_m))^t$ ,  $\mathbf{Y}$  es el vector con las observaciones  $(Y_1, \dots, Y_n)$ , y  $\mathbf{L}$  es la matriz  $m \times n$  definida por la expresión anterior, se tiene que  $\hat{\mathbf{m}} = \mathbf{LY}$ .

v) En particular si queremos evaluar el estimador en los puntos de la muestra  $(X_1, \dots, X_n)$ , tendríamos que calcular la matriz  $n \times n$   $\mathbf{L}$ , donde el elemento  $(i, j)$  viene dado por la expresión

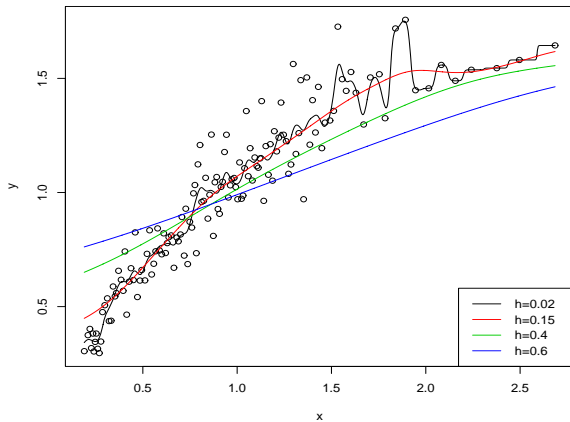
$$L_{ij} = W_j(X_i) = \frac{K\left(\frac{X_i - X_j}{h}\right)}{\sum_{k=1}^n K\left(\frac{X_i - X_k}{h}\right)}, \quad i = 1, \dots, n, j = 1, \dots, n.$$

Una posibilidad sería calcular primero la matriz con los elementos  $K((X_i - X_j)/h)$  y posteriormente dividir cada fila por la suma de la misma. Una vez calculada la matriz  $\mathbf{L}$  se tiene que

$$\hat{\mathbf{m}} = \mathbf{LY}$$

donde  $\hat{\mathbf{m}} = (\hat{m}(X_1), \dots, \hat{m}(X_n))^t$

vi) El parámetro  $h$  controla el grado de suavidad del estimador.





vii) Si  $h \rightarrow 0$  entonces

$$\hat{m}_{n,K}(X_i) \rightarrow Y_i,$$

por tanto, el estimador tiende a interpolar los datos si  $h$  es pequeño (infrasuavizado). Por otra parte, si  $h \rightarrow \infty$  entonces

$$\hat{m}_{n,K}(X_i) \rightarrow \bar{Y},$$

es decir, el estimador es una función constante (sobresuavizado).

### Nota

Puede ocurrir, en zonas donde hay pocos datos, que el denominador de  $W_j(x)$  valga cero. En este caso, como el numerador también valdría cero, se considera que el estimador no está definido.

Una posibilidad para elegir el parámetro de suavizado es usar el método de validación cruzada, convenientemente adaptado al contexto de regresión. Para medir la bondad de ajuste que se consigue con una ventana  $h$  podríamos usar el error medio

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{m}_{n,K}(X_i))^2.$$

Esta medida de error global aproximaría el error de predicción. Sin embargo, la aproximación sería un tanto optimista ya que estaríamos usando el valor de  $Y_i$  dos veces: una a la hora de medir el error, y otra a la hora de construir el estimador.

Para evaluar mejor el error de predicción se suele eliminar el dato  $i$ -ésimo cuando calculamos el error de predicción para  $Y_i$ . Así la función de validación cruzada se define como

$$CV(h) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{m}_{-(i),K}(X_i))^2,$$

donde  $\hat{m}_{-(i),K}$  denota el estimador de Nadaraya-Watson construido a partir de la muestra original después de eliminar el par  $(X_i, Y_i)$ .

La idea sería tomar aquel  $h$  que haga que  $CV$  sea mínimo.

Aunque se podría calcular directamente  $CV$ , esto requeriría evaluar, para cada  $h$ ,  $n$  veces el estimador de Nadaraya-Watson, construido a partir de una muestra de  $(n - 1)$  puntos. Muchos de estos cálculos serían redundantes y se pueden simplificar.

## Teorema

La función de validación cruzada del estimador de Nadaraya-Watson se puede escribir de la siguiente forma

$$CV(h) = \frac{1}{n} \sum_{i=1}^n \left( \frac{Y_i - \hat{m}_{n,K}(X_i)}{1 - L_{ii}} \right)^2$$

donde  $L_{ii}$  es el elemento  $i$ -de la diagonal de la matriz de suavizado  $L$  necesaria para calcular el estimador en los puntos  $(X_1, \dots, X_n)$ .

Es decir

$$L_{ii} = \frac{K(0)}{\sum_{k=1}^n K\left(\frac{X_i - X_k}{h}\right)}.$$

El resultado anterior es un caso particular de un resultado más general que existe para estimadores lineales. Sea

$$\hat{m}(x) = \sum_{j=1}^n l_j(x) Y_j,$$

un estimador lineal. Se define

$$\hat{m}_{(-i)}(x) = \sum_{j=1}^n l_{j,(-i)}(x) Y_j,$$

donde

$$l_{j,(-i)}(x) = \begin{cases} 0 & \text{si } j = i \\ \frac{l_j(x)}{\sum_{k \neq i} l_k(x)} & \text{si } j \neq i. \end{cases}$$

Igual que antes, se define la función de validación cruzada como

$$CV(h) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{m}_{-(i)}(X_i))^2.$$

### Teorema

Se tiene que

$$CV(h) = \frac{1}{n} \sum_{i=1}^n \left( \frac{Y_i - \hat{m}(X_i)}{1 - L_{ii}} \right)^2$$

donde  $L_{ii} = l_i(x_i)$ .

La idea de este método es muy sencilla. En lugar de hacer un ajuste global por mínimos cuadrados de una recta podemos intentar buscar una recta que ajuste bien sólo en los puntos próximos a  $x$ . Dado  $h > 0$  podemos proponer un modelo lineal válido sólo en el entorno  $(x - h, x + h)$

$$Y_i = \alpha(x) + \beta(x)X_i + \epsilon_i, \quad X_i \in (x - h, x + h).$$

Ajustaríamos entonces por mínimos cuadrados los parámetros del modelo usando sólo los datos del entorno local  $(x - h, x + h)$

$$\sum_{i=1}^n (Y_i - \alpha(x) - \beta(x)X_i)^2 \mathbb{I}(|X_i - x| \leq h).$$

Minimizar la suma de cuadrados anterior es equivalente a minimizar

$$\sum_{i=1}^n (Y_i - \alpha(x) - \beta(x)X_i)^2 \omega_h(x - X_i),$$

donde, recordemos,  $\omega$  es la densidad uniforme en  $(-1, 1)$



Al igual que ocurría en la estimación de la densidad, no parece del todo razonable que en la suma de cuadrados anterior tenga el mismo peso todos los errores del intervalo  $(x - h, x + h)$ , independientemente de su proximidad a  $x$ . Para corregir eso podemos reemplazar la suma de cuadrados anterior por

$$\sum_{i=1}^n (Y_i - \alpha(x) - \beta(x)X_i)^2 K_h(x - X_i),$$

donde  $K$  una función de densidad unimodal y simétrica alrededor del cero.

El estimador lineal local en el punto  $x$  vendrá dado por

$$\hat{m}_{n,LL}(x) = a(x) + b(x)x,$$

donde  $a(x), b(x)$  son los valores que minimizan la suma de cuadrados ponderada

$$\sum_{i=1}^n (Y_i - \alpha(x) - \beta(x)X_i)^2 K_h(x - X_i).$$

Por tanto para evaluar el estimador lineal local tenemos que encontrar  $a$  y  $b$  que minimicen

$$\sum_{i=1}^n (Y_i - \alpha - \beta X_i)^2 W_{hi},$$

donde  $W_{hi} = K_h(x - X_i) / \sum_{j=1}^n K_h(x - X_j)$ . Si derivamos con respecto a las variables  $\alpha$  y  $\beta$  ( $x$  está fijo) obtenemos la ecuaciones (ejercicio)

$$\begin{aligned} \sum_{i=1}^n W_{hi} Y_i &= a + b \sum_{i=1}^n W_{hi} X_i \\ \sum_{i=1}^n W_{hi} X_i Y_i &= a \left( \sum_{i=1}^n W_{hi} X_i \right) + b \left( \sum_{i=1}^n W_{hi} X_i^2 \right) \end{aligned}$$

Si utilizamos la notación  $\mu_{r,s}^h = \sum_{i=1}^n W_{hi} X_i^r Y_i^s$  el sistema anterior puede escribirse como

$$\begin{aligned}\mu_{0,1}^h &= a + b\mu_{1,0}^h \\ \mu_{1,1}^h &= a\mu_{1,0}^h + b\mu_{2,0}^h\end{aligned}$$

Despejando (ejercicio)

$$a = \mu_{0,1}^h - b\mu_{1,0}^h, \quad b = \frac{\mu_{1,1}^h - \mu_{1,0}^h\mu_{0,1}^h}{\mu_{2,0}^h - (\mu_{1,0}^h)^2}$$

¿A qué convergen  $a$  y  $b$  si  $h \rightarrow \infty$ ?

Al igual que ocurría con el estimador de Nadaraya-Watson, el estimador lineal local también es un estimador lineal. Por tanto, la función de validación cruzada es bastante sencilla de calcular.

## Teorema

El estimador local lineal se puede escribir de la forma

$$\hat{m}_{n,LL}(x) = \sum_{j=1}^n l_j(x) Y_j, \text{ donde } l_j(x) = \frac{b_j(x)}{\sum_{k=1}^n b_k(x)},$$

con

$$b_j(x) = K \left( \frac{x_j - x}{h} \right) (S_{n,2}(x) - (x_j - x) S_{n,1}(x)),$$

y

$$S_{n,r}(x) = \sum_{j=1}^n K \left( \frac{x_j - x}{h} \right) (x_j - x)^r, \quad r = 1, 2.$$

Supongamos que en vez de haber ajustado una recta localmente se hubiera ajustado una constante en el entorno  $(x - h, x + h)$ . Se trataría de buscar aquel  $a$  que minimice la suma de cuadrados ponderada

$$\sum_{i=1}^n (Y_i - \alpha)^2 K_h(x - X_i),$$

o, equivalentemente,

$$\sum_{i=1}^n (Y_i - \alpha)^2 W_{hi}.$$

Por las propiedades de la media se sabe que (ejercicio)

$$a = \sum_{i=1}^n W_{hi} Y_i = \hat{m}_{n,K}(x),$$

es decir el estimador de Nadaraya-Watson en  $x$ .

En la siguiente tabla se muestra el sesgo y varianza asintótica (para diseño aleatorio) de los estimadores de Nadaraya-Watson y local lineal:

Método	Sesgo	Varianza
Nadaraya-Watson	$(m''(x) + \frac{2m'(x)f'(x)}{f(x)})b_n$	$V_n$
Local lineal	$m''(x)b_n$	$V_n$

donde

$$b_n = \frac{1}{2}\mu_2(K)h^2, \quad V_n = \frac{\sigma^2(x)}{f(x)nh}R(K)$$

Es sorprendente que a pesar de estimar un parámetro más en el estimador local lineal que en el de Nadaraya-Watson la varianza asintótica es la misma.

Es posible generalizar el método local lineal y ajustar localmente un polinomio de grado  $p$ . Supongamos que  $m$  tiene  $(p+1)$  derivadas continuas en un entorno de  $x$ . Por el teorema de Taylor

$$m(z) \approx m(x) + m'(x)(z-x) + \frac{m''(x)}{2!}(z-x)^2 + \cdots + \frac{m^{(p)}(x)}{p!}(z-x)^p.$$

Podemos ajustar localmente este polinomio de grado  $p$  mediante mínimos cuadrados ponderados. Habría que encontrar el parámetro  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)^t$  que minimiza la función

$$\Psi(\beta) = \sum_{i=1}^n \left( Y_i - \sum_{j=0}^p \beta_j (X_i - x)^j \right)^2 K_h(x - X_i)$$



Nótese que  $\hat{\beta}_j$  estima

$$\beta_j = \frac{f^{(j)}(x)}{j!}, \quad j = 0, \dots, p.$$

En particular

$$\hat{m}_{n,PL}(x) = \hat{\beta}_0,$$

es el estimador polinómico local de orden  $p$  de  $m(x)$ .

El problema anterior se puede escribir de forma matricial. Sean

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x) & \cdots & (X_1 - x)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (X_n - x) & \cdots & (X_n - x)^p \end{pmatrix}$$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\mathbf{W}_X = \begin{pmatrix} K_h(x - X_1) & 0 & \cdots & 0 \\ 0 & K_h(x - X_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & K_h(x - X_n) \end{pmatrix}$$

Con esta notación la función  $\Psi$  puede escribirse así

$$\Psi(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^t \mathbf{W} (\mathbf{Y} - \mathbf{X}\beta).$$

Es conocido que su mínimo se alcanza en

$$\hat{\beta}(x) = (\mathbf{X}^t \mathbf{W}_x \mathbf{X})^{-1} \mathbf{X}^t \mathbf{W}_x \mathbf{Y} = \mathbf{H} \mathbf{Y}$$

El vector solución tiene  $(p + 1)$  componentes. La componente  $j$ -ésima permite estimar la derivada correspondiente de  $m$

$$\hat{m}^{(j)}(x) = j! \hat{\beta}_j(x).$$

En particular

$$\hat{m}_{n,PL}(x) = \mathbf{e}_1^t \mathbf{H} \mathbf{Y},$$

siendo  $\mathbf{e}_1 = (1, 0, \dots, 0)^t \in \mathbb{R}^{p+1}$ . Por tanto,

$$\hat{m}_{n,PL}(x) = \sum_{j=1}^n l_j(x) Y_j,$$

es un estimador lineal, donde  $l(x) = (l_1(x), \dots, l_n(x))^t$  viene dado por la expresión

$$l(x) = \mathbf{e}_1^t (\mathbf{X}^t \mathbf{W}_x \mathbf{X})^{-1} \mathbf{X}^t \mathbf{W}_x.$$