Estimación No Paramétrica de la Regresión

IECD

Ejemplo

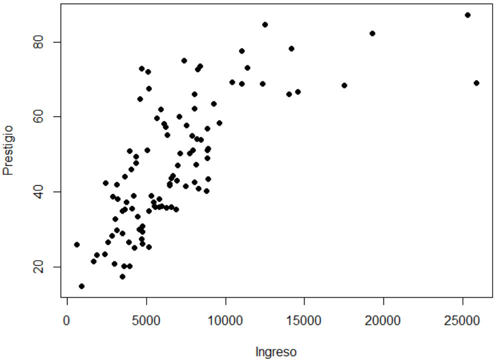
Los datos, tomados en Canadá en 1970, corresponden a 102 ocupaciones y se registraron, entre otras, las variables:

* ingreso: en dólares
* prestigio: de 0 a 100.

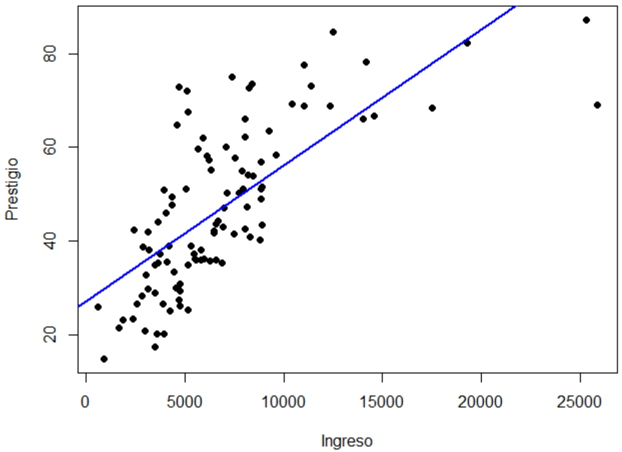
Ejemplo

Los datos, tomados en Canadá en 1970, corresponden a 102 ocupaciones y se registraron, entre otras, las variables:

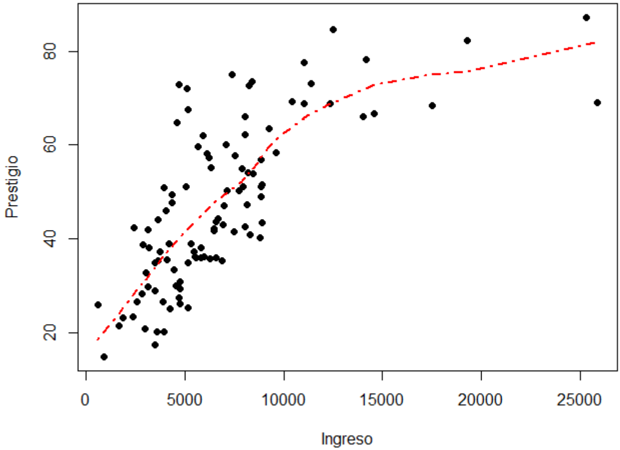
* ingreso: en dólares
* prestigio: de 0 a 100.



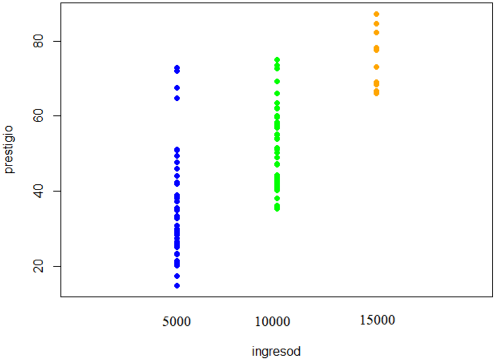
Y =prestigio vs. X = ingreso: recta de mínimos cuadrados



Y vs. X: ¿Cómo captar esta tendencia que vemos?



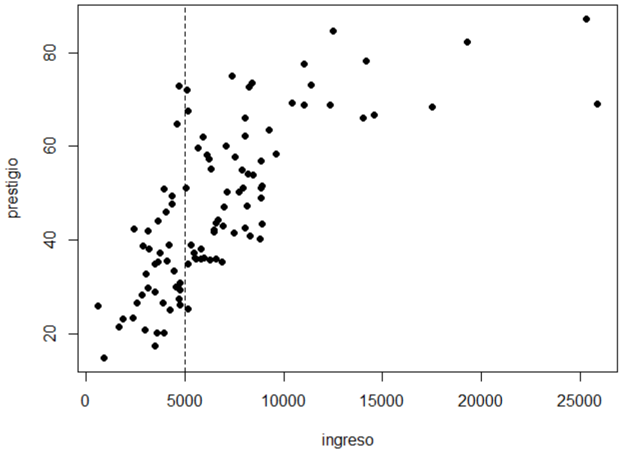
X categórica: si ingreso fuera discreta...



La pregunta:

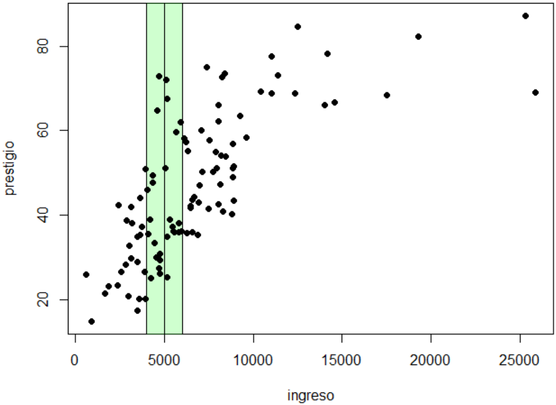
¿Qué prestigio tendrá una ocupación con un ingreso igual a 5000?

X continua



¿Qué prestigio tendrá una ocupación con un ingreso igual a 5000?

Abrimos una ventana



Promediamos los valores de prestigio de los que están dentro de la ventana.

Propuesta inicial: Predecir con promedios locales

1. Determinamos el valor de x donde se quiere predecir (5000).

2. Elegimos un valor h de ventana para armar la vecindad.

3. Promediamos los valores de la respuesta correspondientes a los pares que caen dentro de la vecindad con ventana de tamaño h centrada en x ($x \pm h$).

Vayamos al shiny a experimentar un poco:

`https://glmconr2.shinyapps.io/app_regre2/`

Estimación no paramétrica de la regresión

Y: respuesta
X: covariable

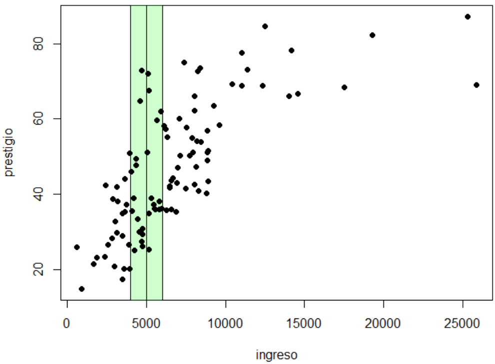
- Modelo $\quad Y = m(X) + \varepsilon$
  X y ε independientes, $\mathbb{E}(\varepsilon) = 0$
- Función de regresión:

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

- Muestra: $\{(X_i, Y_i) : i = 1, \ldots, n\}$

X: discreta $\qquad \widehat{m}(x) = \dfrac{\displaystyle\sum_{i=1}^{n} Y_i \; I_{\{X_i = x\}}}{\displaystyle\sum_{j=1}^{n} I_{\{X_j = x\}}}$

X continua: ventana h



Estimador de Núcleos de Nadaraya - Watson

- Estimación de $m(x) = \mathbb{E}(Y \mid X = x)$ - X continua.

$$\widehat{m}_h(x) = \frac{\displaystyle\sum_{i=1}^{n} Y_i I_{\{|X_i - x| \le h\}}}{\displaystyle\sum_{j=1}^{n} I_{\{|X_j - x| \le h\}}}$$

$$\widehat{m}_h(x) = \frac{\displaystyle\sum_{i=1}^{n} Y_i I_{\left\{\left|\frac{X_i - x}{h}\right| \le 1\right\}}}{\displaystyle\sum_{j=1}^{n} I_{\left\{\left|\frac{X_j - x}{h}\right| \le 1\right\}}}$$

Estimador de Núcleos de Nadaraya - Watson

- Estimación de $m(x) = \mathbb{E}(Y \mid X = x)$ - X continua.

$$\widehat{m}_h(x) = \frac{\displaystyle\sum_{i=1}^{n} Y_i I_{\{|X_i - x| \leq h\}}}{\displaystyle\sum_{j=1}^{n} I_{\{|X_j - x| \leq h\}}}$$

$$\widehat{m}_h(x) = \frac{1/2 \displaystyle\sum_{i=1}^{n} Y_i I_{\left\{\left|\frac{X_i - x}{h}\right| \leq 1\right\}}}{1/2 \displaystyle\sum_{j=1}^{n} I_{\left\{\left|\frac{X_j - x}{h}\right| \leq 1\right\}}}$$

Estimador de Núcleos de Nadaraya - Watson

- Estimación de $m(x) = \mathbb{E}(Y \mid X = x)$ - X continua.

$$\widehat{m}_h(x) = \frac{1/2 \displaystyle\sum_{i=1}^{n} Y_i I_{\left\{\left|\frac{X_i - x}{h}\right| \leq 1\right\}}}{1/2 \displaystyle\sum_{j=1}^{n} I_{\left\{\left|\frac{X_j - x}{h}\right| \leq 1\right\}}}$$

$$K(u) = \frac{1}{2} I_{\{|u| \leq 1\}}\ , K = f_U, U \sim \mathcal{U}[-1, 1]$$

$$\widehat{m}_h(x) = \frac{\displaystyle\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)}{\displaystyle\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}$$

Estimador de Núcleos de Nadaraya–Watson

$$
\begin{aligned}
\widehat{m}_h(x) &= \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)} \\
&= \sum_{i=1}^n Y_i \; \underbrace{\frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)}}_{= \, W_{i,h}(x)} \\
\widehat{m}_h(x) &= \sum_{i=1}^n Y_i \, W_{i,h}(x) \\
&= \text{promedio ponderado por la distancia a } x.
\end{aligned}
$$

Tipos de núcleos

- Núcleo Rectangular: $K(t) = \frac{1}{2}\mathcal{I}_{[-1,1]}(t)$
- Núcleo Triangular: $K(t) = (1 - |t|)\mathcal{I}_{[-1,1]}(t)$
- Núcleo Gausssiano: $K(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}$
- Núcleo Epanechnikov: $K(t) = \frac{3}{4}(1 - t^2)\mathcal{I}_{[-1,1]}(t)$

Estimador de Nadaraya–Watson

Proponemos estimadores de la forma:

$$\widehat{m}_h(x) = \sum_{i=1}^{n} Y_i \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}$$

$K(t)$ es un kernel y $h > 0$ es el ancho de ventana, donde $K(t)$ satisface:

a) $K(t) \geq 0$

b) $K(t) = K(-t)$ (función par)

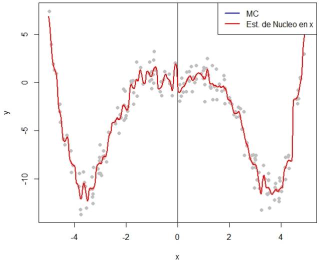
c) $\int K(t)dt = 1$

d) $\int tK(t)dt = 0$

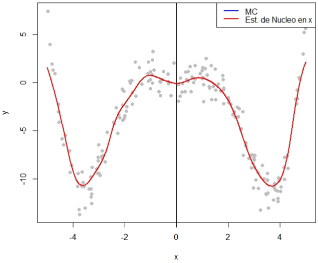
e) $\int t^2 K(t)dt < \infty$

¡Nosotros elegimos a K!!

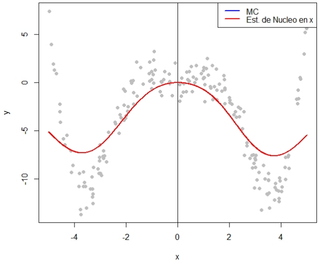
h=0.05



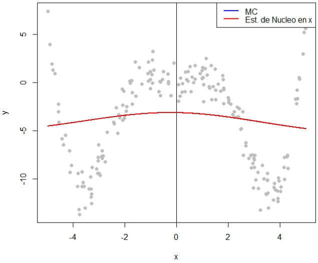
h=0.30



h=1



h=3



Estimador de Nadaraya–Watson

$$\widehat{m}_h(x) = \sum_{i=1}^{n} Y_i \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}$$

A1. m es dos veces diferenciable

A2. f, la densidad de X, es C^1 y acotada lejos del 0

A3. K cumple las condiciones a) a e).

A4. $h \to 0$ cuando $n \to \infty$ y además $nh \to \infty$

Estimador de Nadaraya–Watson

$$\widehat{m}_h(x) = \sum_{i=1}^{n} Y_i \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}$$

A1. m es dos veces diferenciable

A2. f, la densidad de X, es C^1 y acotada lejos del 0

A3. K cumple las condiciones a) a e).

A4. $h \to 0$ cuando $n \to \infty$ y además $nh \to \infty$

Teorema: Bajo las condiciones A1 a A4 tenemos que

$$\mathrm{B}\left[\widehat{m}_h(x)\right] = \frac{h^2}{2}\mu_2(K)\left\{m''(x) + 2\frac{m'(x)f'(x)}{f(x)}\right\} + o\left(h^2\right)$$

$$\mathbb{V}ar\left[\widehat{m}_h(x)\right] = \frac{1}{nh}\frac{\sigma^2\mathbb{C}(K)}{f(x)} + o\left((nh)^{-1}\right)$$

donde

$\mu_2(K) = \int t^2 K(t)dt$, $\mathbb{C}(K) = \int K^2(t)dt$ y $\mathbb{V}ar(\varepsilon) = \sigma^2$.

$$\mathrm{B}\left[\widehat{m}_h(x)\right] = \frac{h^2}{2}\mu_2(K)\left\{m''(x) + 2\frac{m'(x)f'(x)}{f(x)}\right\} + o\left(h^2\right)$$

$$\mathbb{V}ar\left[\widehat{m}_h(x)\right] = \frac{1}{nh}\frac{\sigma^2\mathbb{C}(K)}{f(x)} + o\left((nh)^{-1}\right)$$

- El sesgo decrece cuadráticamente con h:
  h pequeña da estimadores con sesgo bajo.
- El sesgo depende de $m''(x)$, la curvatura de m en x:
  - negativamente cuando $m''(x) < 0$ (picos y modas de m)
  - positivamente cuando $m''(x) > 0$ (valles de m)
  - en términos absolutos, a mayor curvatura, más sesgo.

$$\mathrm{B}\left[\widehat{m}_h(x)\right] = \frac{h^2}{2}\mu_2(K)\left\{m''(x) + 2\frac{m'(x)f'(x)}{f(x)}\right\} + o\left(h^2\right)$$

$$\mathbb{V}ar\left[\widehat{m}_h(x)\right] = \frac{1}{nh}\frac{\sigma^2\mathbb{C}(K)}{f(x)} + o\left((nh)^{-1}\right)$$

- El sesgo decrece cuadráticamente con h:
  h pequeña da estimadores con sesgo bajo.
- El sesgo depende de $m''(x)$, la curvatura de m en x:
  - negativamente cuando $m''(x) < 0$ (picos y modas de m)
  - positivamente cuando $m''(x) > 0$ (valles de m)
  - en términos absolutos, a mayor curvatura, más sesgo.
- Tanto sesgo como varianza aumentan cuanto más pequeña sea $f(x)$.

$$\mathrm{B}\left[\widehat{m}_h(x)\right] = \frac{h^2}{2}\mu_2(K)\left\{m''(x) + 2\frac{m'(x)f'(x)}{f(x)}\right\} + o\left(h^2\right)$$

$$\mathbb{V}ar\left[\widehat{m}_h(x)\right] = \frac{1}{nh}\frac{\sigma^2\mathbb{C}(K)}{f(x)} + o\left((nh)^{-1}\right)$$

- El sesgo decrece cuadráticamente con h:
  h pequeña da estimadores con sesgo bajo.
- El sesgo depende de $m''(x)$, la curvatura de m en x:
  - negativamente cuando $m''(x) < 0$ (picos y modas de m)
  - positivamente cuando $m''(x) > 0$ (valles de m)
  - en términos absolutos, a mayor curvatura, más sesgo.
- Tanto sesgo como varianza aumentan cuanto más pequeña sea $f(x)$.
- La varianza disminuye a medida que nh crece.
  (nh:*tamaño efectivo de muestra*, cantidad de datos en la vecindad que son usados para la estimación de m en x).
- Para disminuir la varianza necesitamos h grande fijado n.

Compromiso Sesgo–Varianza $\rightarrow MSE$

Error Cuadrático Medio Integrado: un camino para elegir h

Para tener una medida del error global que cometemos podemos considerar el **Error Cuadrático Medio Integrado**

$$\begin{aligned} MISE &= \mathbb{E}\left(\int [\widehat{m}_h(x) - m(x)]^2 f(x)dx\right) \\ &= \int \mathbb{E}[\widehat{m}_h(x) - m(x)]^2 f(x)dx \\ &= \int MSE(\widehat{m}_h(x)) f(x)dx \end{aligned}$$

Por las expresiones vistas, tenemos para constantes $\mathbb{C}_i$ apropiadas

$$MISE \approx \mathbb{C}_1 h^4 + \frac{\mathbb{C}_2}{nh}$$

Esto nos da un camino para elegir: minimizar los términos dominantes del $MISE$

$$h = \left(\frac{\mathbb{C}_2}{4\ \mathbb{C}_1 n}\right)^{1/5}$$

Otro camino para elegir h: Convalidación Cruzada

$\mathcal{M} : (X_1, Y_1), \ldots, (X_n, Y_n)$ independientes $\to \widehat{m}_h$

Dada una nueva observación (X_o, Y_o) definimos

Error o Riesgo de Predicción: $R(h) = \mathbb{E}[(Y_o - \widehat{m}_h(X_o))^2 | \mathcal{M}]$

Aproximar este error por la suma de cuadrados residual

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}_h(X_i))^2$$

no sería buena idea: peligro de sobreajuste.

Otro camino para elegir h: Convalidación Cruzada

Error o Riesgo de Predicción: $R(h) = \mathbb{E}[(Y_o - \widehat{m}_h(X_o))^2 | \mathcal{M}]$

Aproximaremos a $R(h)$ usando el enfoque de predicción cruzada dividiendo los datos.

Propuesta: considerar el método *dejando–uno–afuera o one–leave–out* y computar el error cuadrático de convalidación cruzada:

$$\mathrm{CV}(h) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{m}_h^{-i}(X_i))^2$$

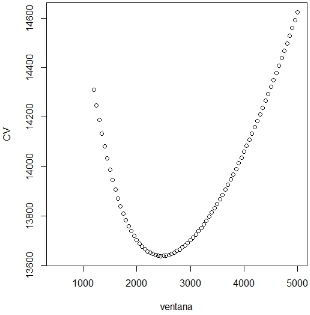
donde $\widehat{m}_h^{-i}(\cdot)$ se computa sin la observación (X_i, Y_i).

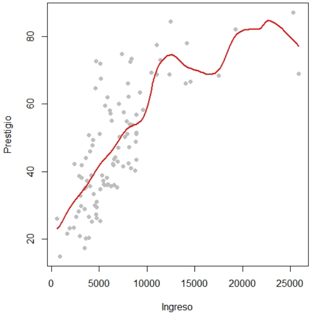
$$h_{CV} = \mathrm{argmin}_{h>0}\, \mathrm{CV}(h)$$

Volviendo al ejemplo inicial

* ingreso: en dólares
* prestigio: índice, de 0 a 100.

Estimador de Nadaraya-Watson con ventana de Convalidación Cruzada





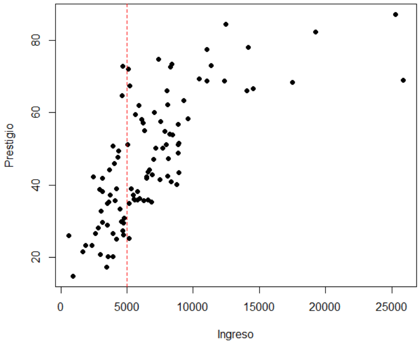
$$h_{CV} = 2450$$

El jardín de senderos que se bifurcan...

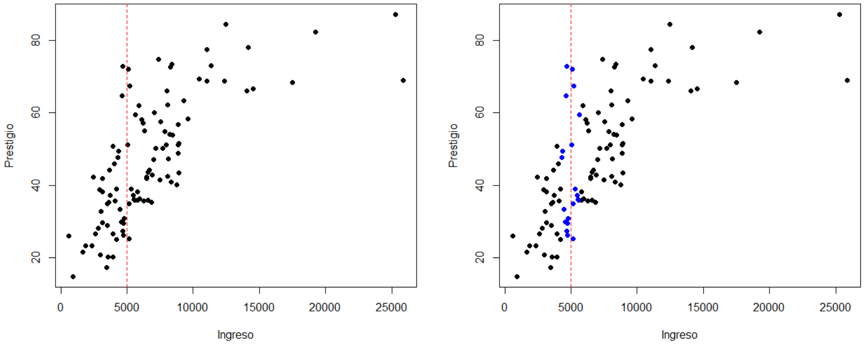
- Elegir los k ingresos más cercanos y promediar las respuestas correspondientes.

- Calcular medianas en lugar de promedios: medianas locales

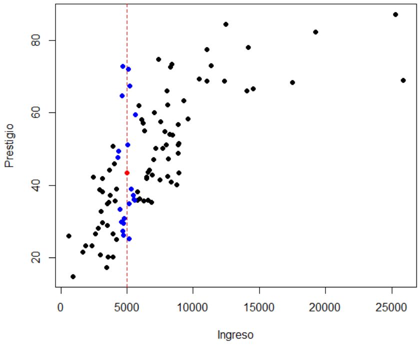
k Vecinos Más Cercanos



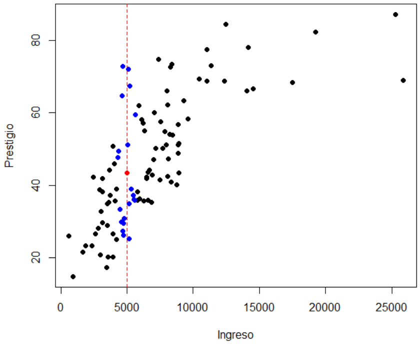
k Vecinos Más Cercanos

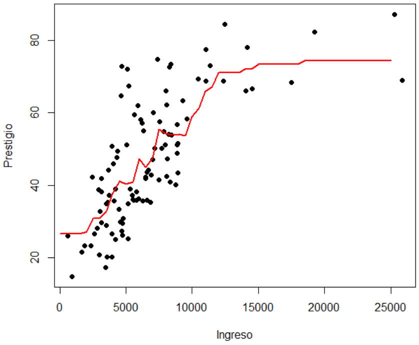


k Vecinos Más Cercanos



k Vecinos Más Cercanos





$k-$vecinos más cercanos (knn: k-nearest neighbours)

El método de $k-$Vecinos más cercanos permite estimar la regresión.

- Elegimos k un entero positivo y un punto x donde predecir.
- Identificamos los índices de los k puntos más cercanos a x. Sea N_x dicho conjunto.
- Estimamos a $m(x)$ por el promedio de los valores de la respuesta en N_x:

$$\widehat{m}(x) = \frac{1}{k} \sum_{i \in N_x} y_i$$

¿Cómo elegimos el parámetro k?

¿Por qué usaríamos medianas?