MiDaS Millenium Nucleus Center for the Discovery of Structures in Complex Data

FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

Agencia Nacional de Investigación y Desarrollo
Ministerio de Ciencia, Tecnología, Conocimiento e Innovación
Gobierno de Chile

milenio INICIATIVA CIENTÍFICA

Asociación entre variables

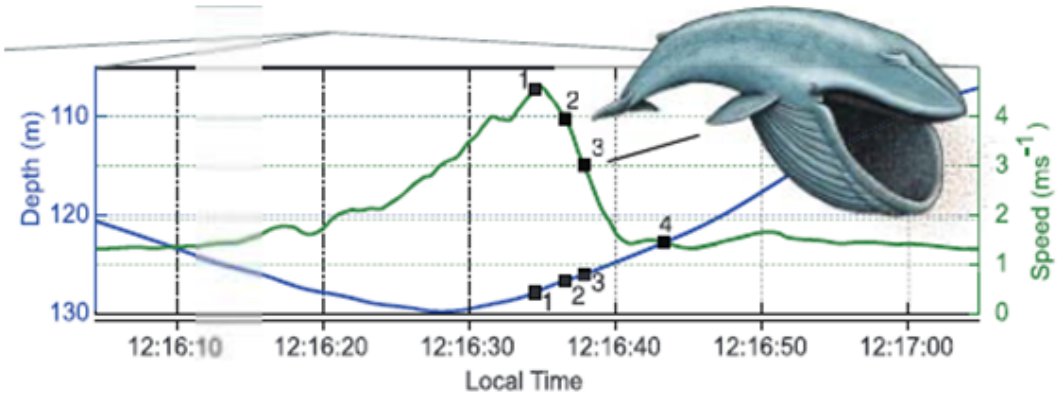
Ana María Araneda

Olimpiada del Big Data

¿Por qué las ballenas son gigantes pero no tan gigantes?

Estudio de Goldbigen et al. (2019)

Las ballenas son etiquetadas utilizando varillas de fibra de carbono y ventosas:



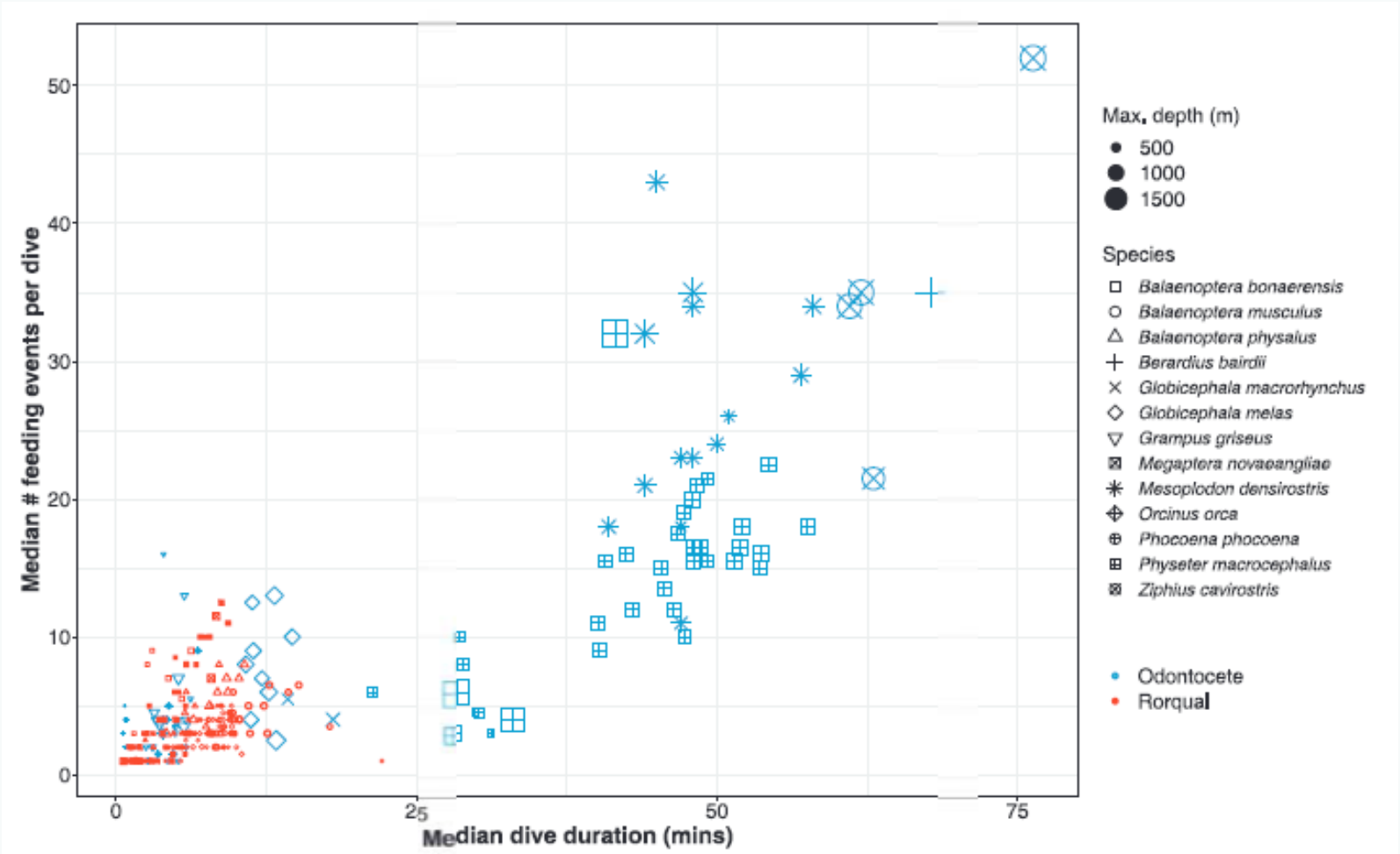
Estudio reportado en revista Science

- Las ballenas pueden clasificarse en *Odontocetes* y *Rorqual*. Cada uno de estos linajes está determinado por su estrategia de alimentación.
- Aunque las ballenas *rorqual* son de mayor tamaño, ambos linajes pueden ser considerados “muy grandes”.
- La bf hipótesis de los investigadores es que la eficiencia energética de su alimentación es mayor en la medida en que aumenta el tamaño de la ballena.
- **El estudio incorpora análisis de múltiples variables que afectan la eficiencia energética de las ballenas.**

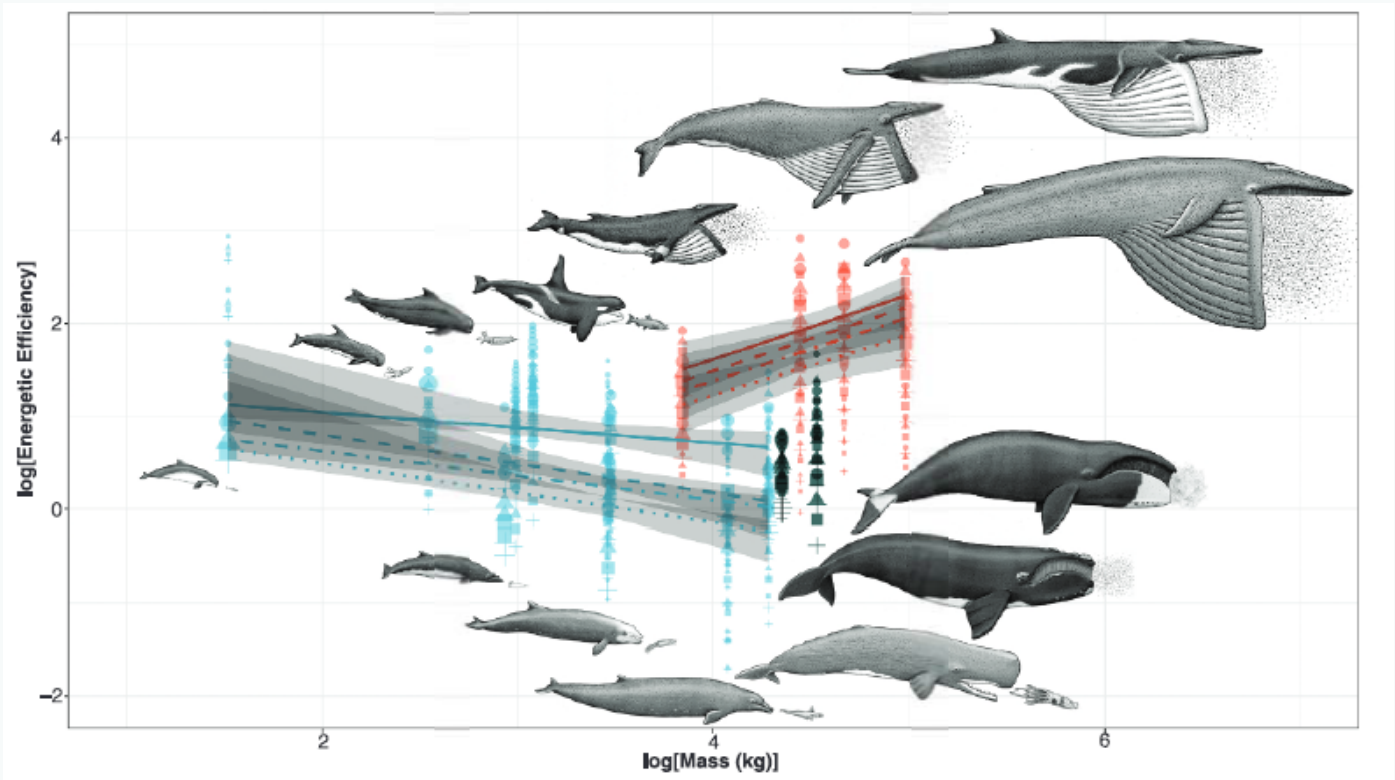
MiDaS

Asociación entre múltiples variables

Eventos de alimentación vs. tiempo de buceo



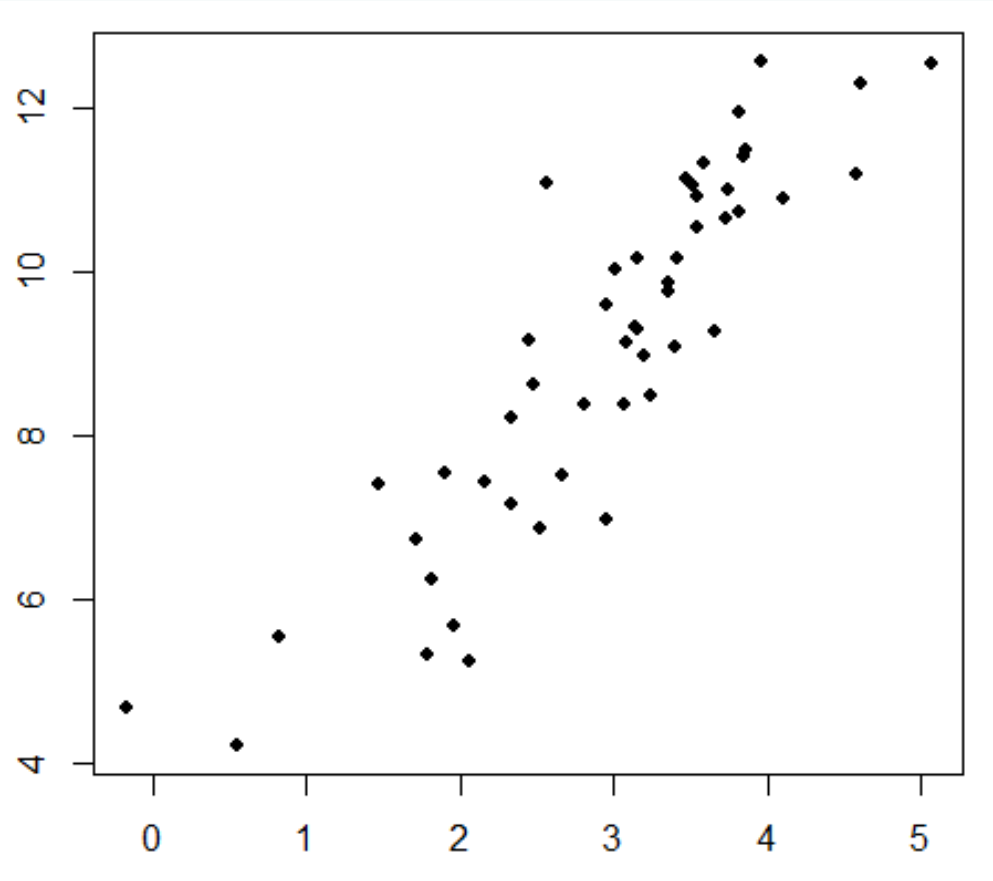
Energía neta vs. masa corporal



Necesidad de cuantificar el nivel de asociación

Coeficiente de correlación lineal de Pearson

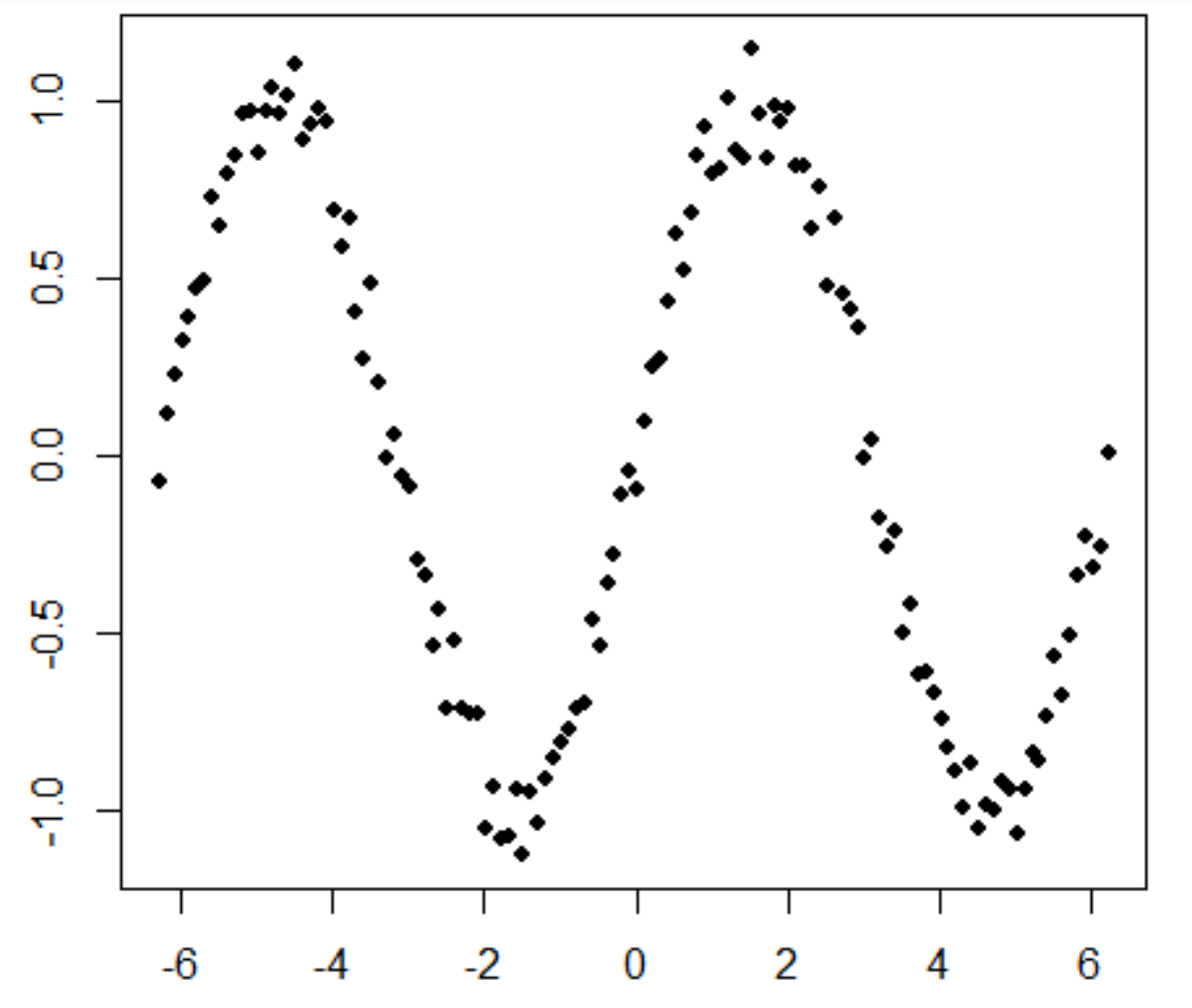
¿Qué mide este coeficiente?



$r = 0,87$

MiDaS

¿Qué no mide este coeficiente?



$$r = -0,25$$

Interpretación de su valor:

MiDaS

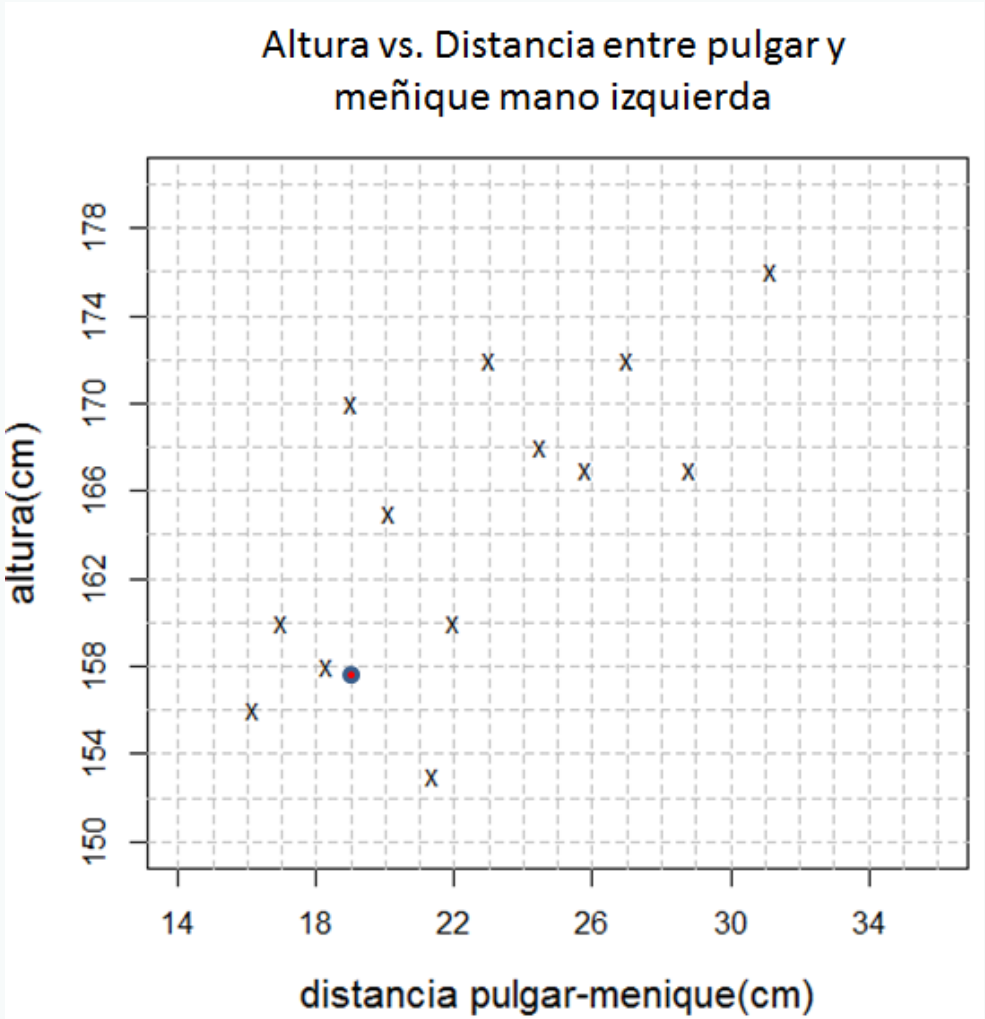
Recta de Mínimos Cuadrados

La *mejor* recta

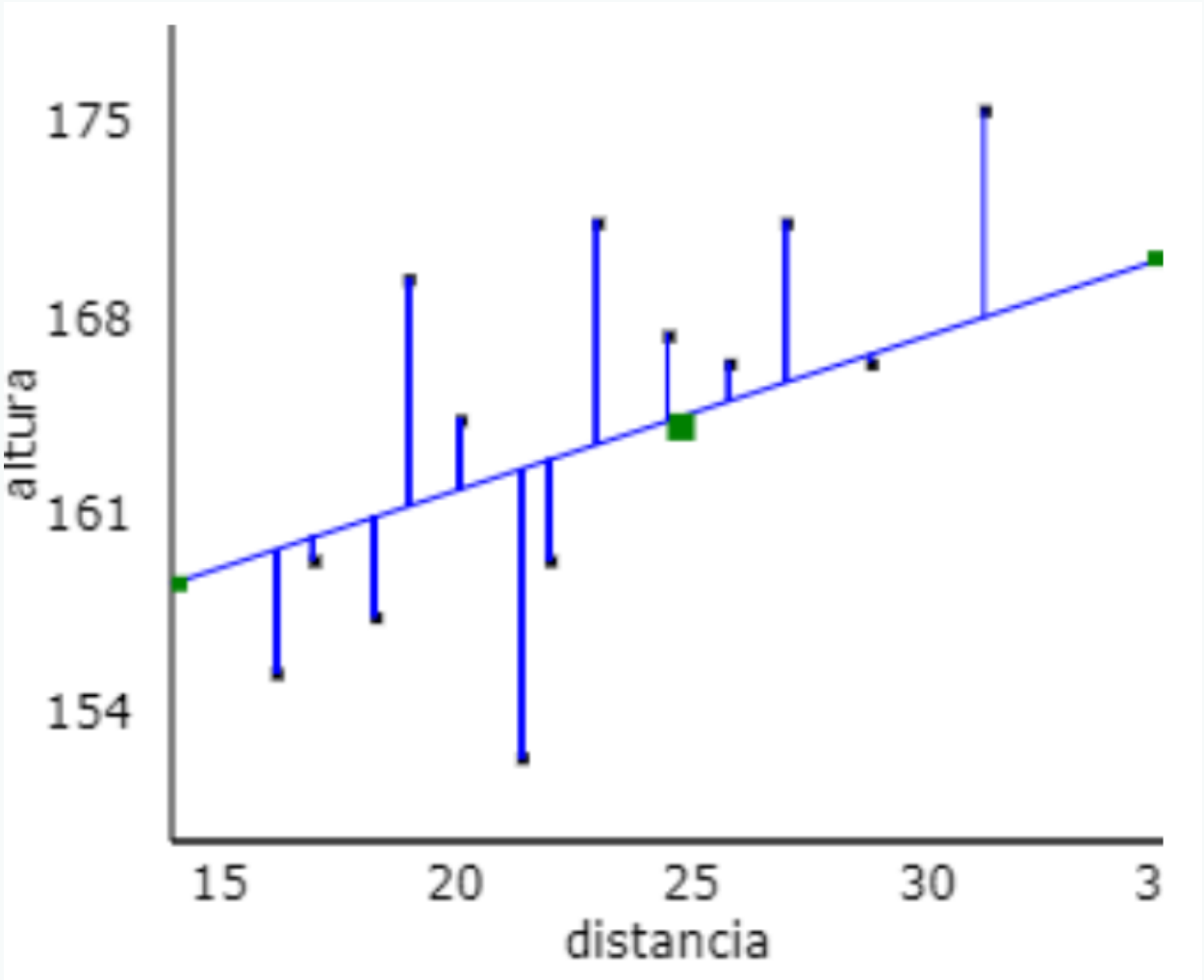
Existe la conjetura de que la altura de los individuos está asociada a otras medidas de su cuerpo, como algunas medidas sobre manos y pies. Esta idea es utilizada en la búsqueda de sujetos sospechosos cuando se dispone de pisadas o impresiones de manos. ¿Existirá en particular, alguna relación entre la altura de una persona y la distancia entre sus dedos pulgar y meñique?

MiDaS

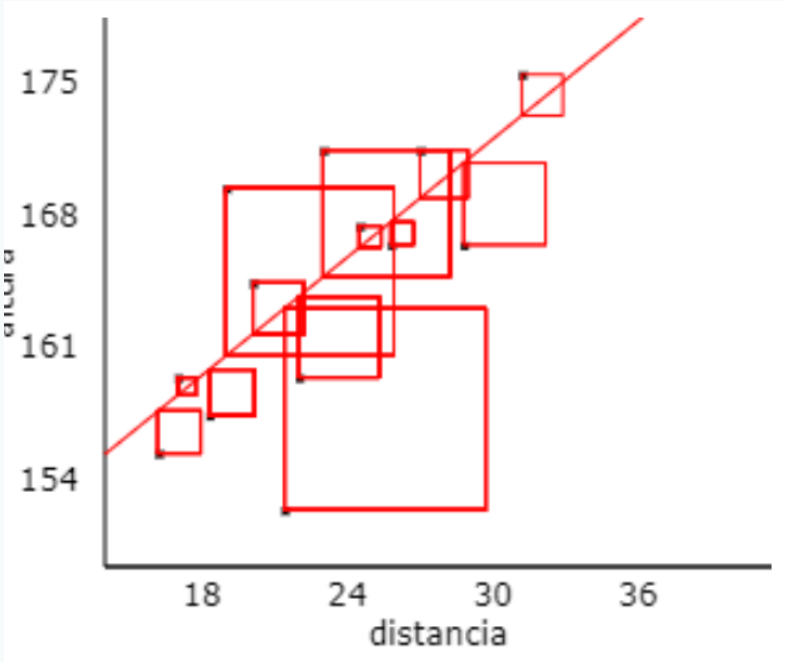
Datos recolectados en curso Estadística y Probabilidad en la UC



Residuos de una recta:



Residuos al cuadrado:

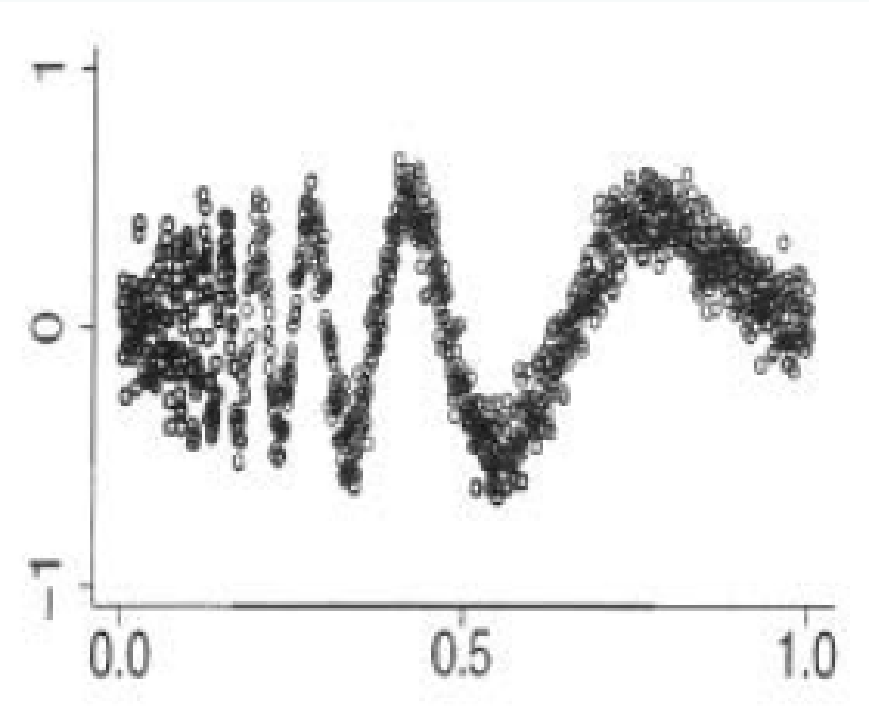


Recta de Mínimos Cuadrados:

Corresponde a la recta que minimiza la suma de las áreas de los cuadrados de l
Es posible obtener su intercepto y pendiente de manera analítica.

MIDaS

Qué pasa con otro tipo de relaciones?



MiDaS

Regresión no paramétrica

Idea:

La recta que estima la relación entre las observaciones depende de parámetros que la hacen poco flexible. Necesitamos estrategias flexibles para modelar relaciones no lineales arbitrarias.

Familia de suavizadores lineales:

En esta familia de modelos, el valor de la curva estimada en un valor del eje de las abscisas corresponde a un promedio ponderado de los valores de la variable respuesta, representada en el eje de las ordenadas.

$$\hat{Y}_i = \sum_{j=1}^{n} \omega_j \, x_j.$$

MiDaS

Algunos suavizadores lineales

Regresograma:

Suponga que para ciertos valores a y b se cumple que:

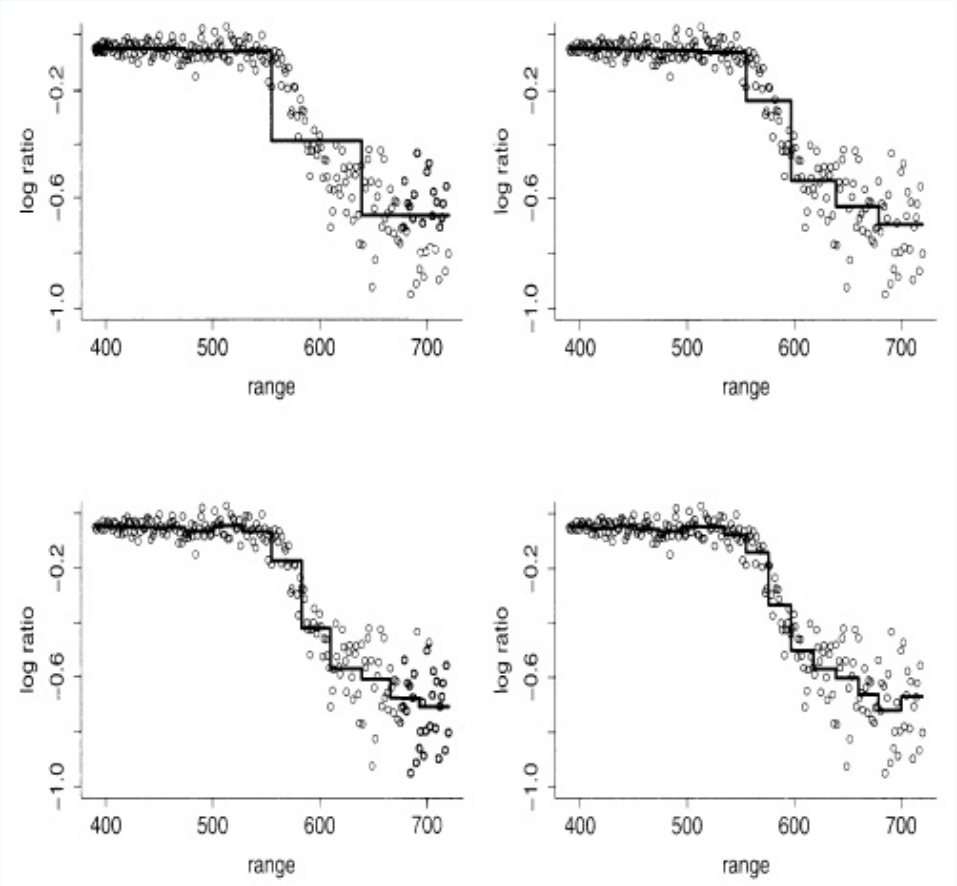
$$a \leq x_i \leq b,$$

para todo $i = 1, \ldots, n$. Se divide el intervalo $[a,\ b]$ en m subintervalos de igual largo.

Para cada valor en el eje de las abscisas, se estima la curva como el promedio de las observaciones que se encuentran en su mismo interval

MiDaS

Efecto del ancho de los subintervalos:



Note que, aunque la figura depende del ancho de los subintervalos definidos, no podría, en general, llamarse *curva suave*.

Promedios Locales:

Sea $h > 0$ un valor constante y sea x un valor fijo en el eje de las abscisas. Definimos una vecindad de x a un intervalo centrado en x de ancho $2h$.

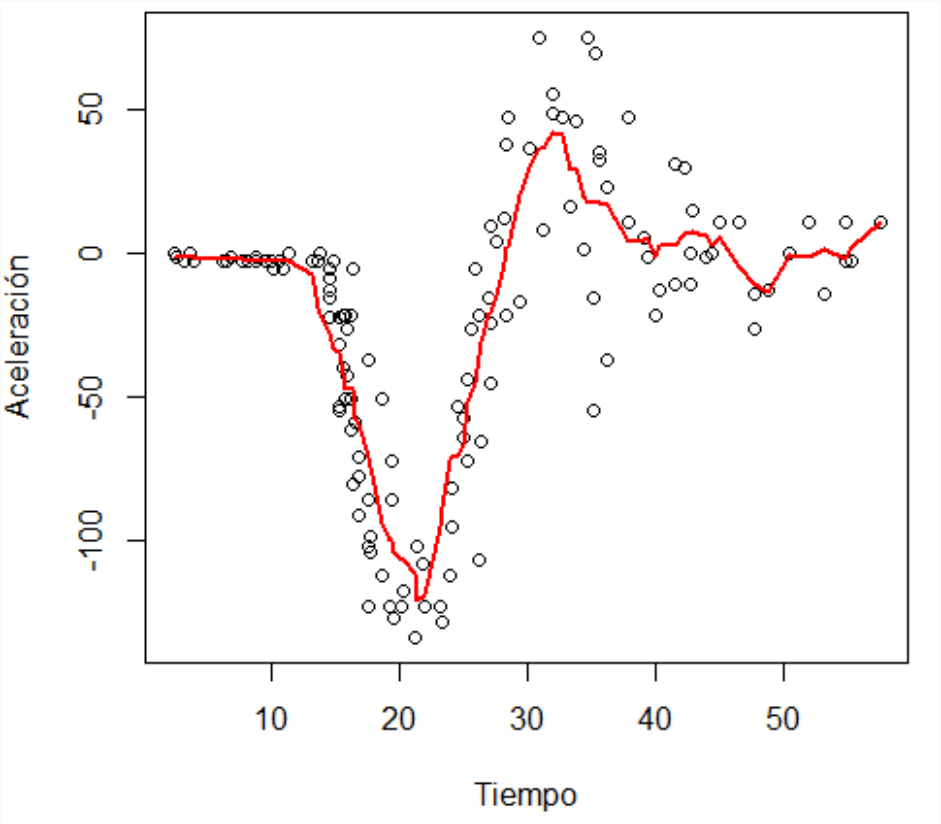
Para cada valor en el eje de las abscisas, se estima la curva como el promedio de las observaciones que se encuentran en su vecindad.

Note que ahora la ventana es móvil.

MiDaS

Ejemplo:

La siguiente figura muestra la curva estimada para datos de aceleración versus tiempo, utilizando promedios locales con $h = 2$.



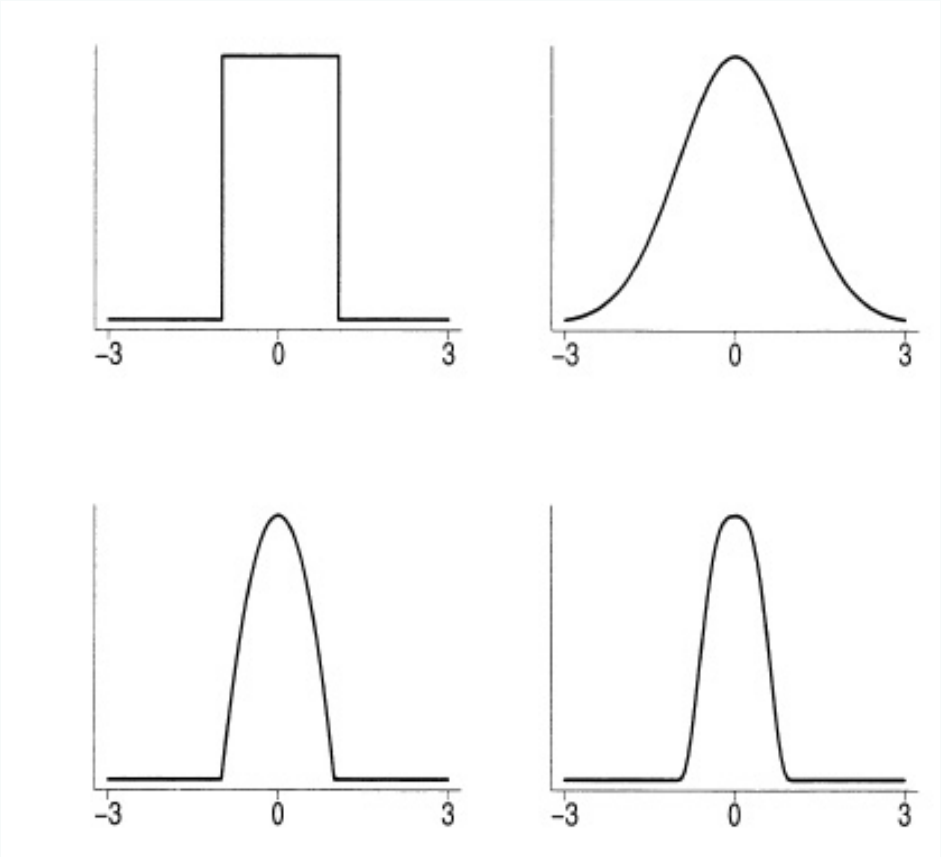
Estimador de Nadaraya-Watson:

La idea es parecida a la de utilizar promedios locales, sin embargo, para cada valor en el eje de las abscisas, el valor estimado de la curva da mayor peso a las observaciones más cercanas. Los pesos van decre

Los pesos quedan determinanos por una función K, que se denomina Kernel.

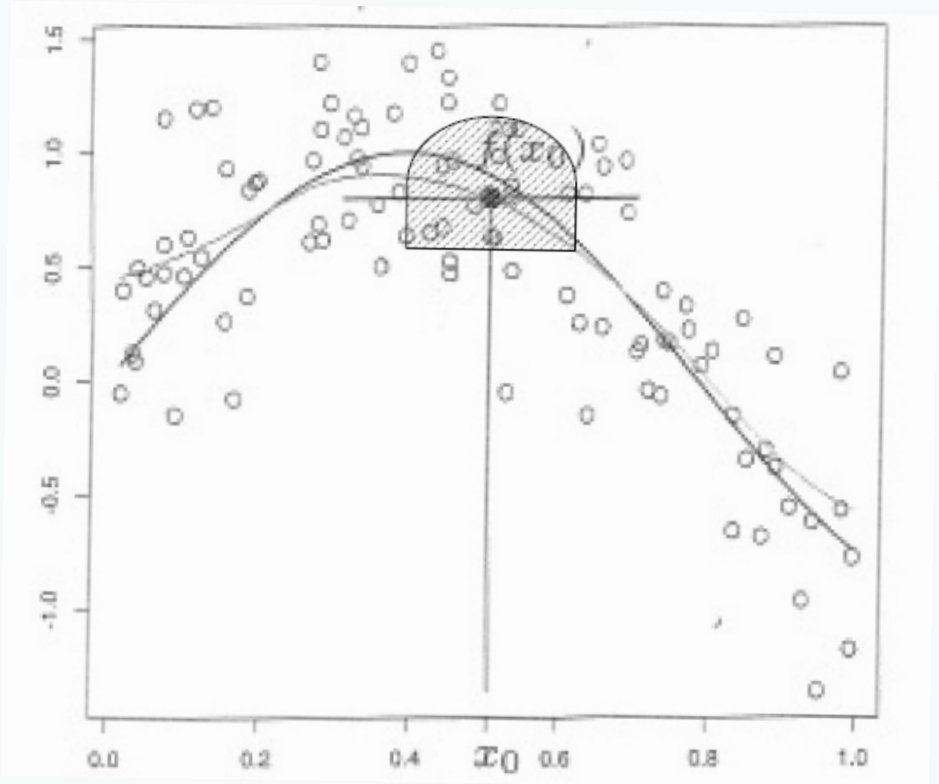
MiDaS

Algunos kernels:



Arriba izquierda: Boxcar. Arriba derecha: Gaussiano. Abajo izquierda: Epanechnikov. Abajo derecha: tricubo.

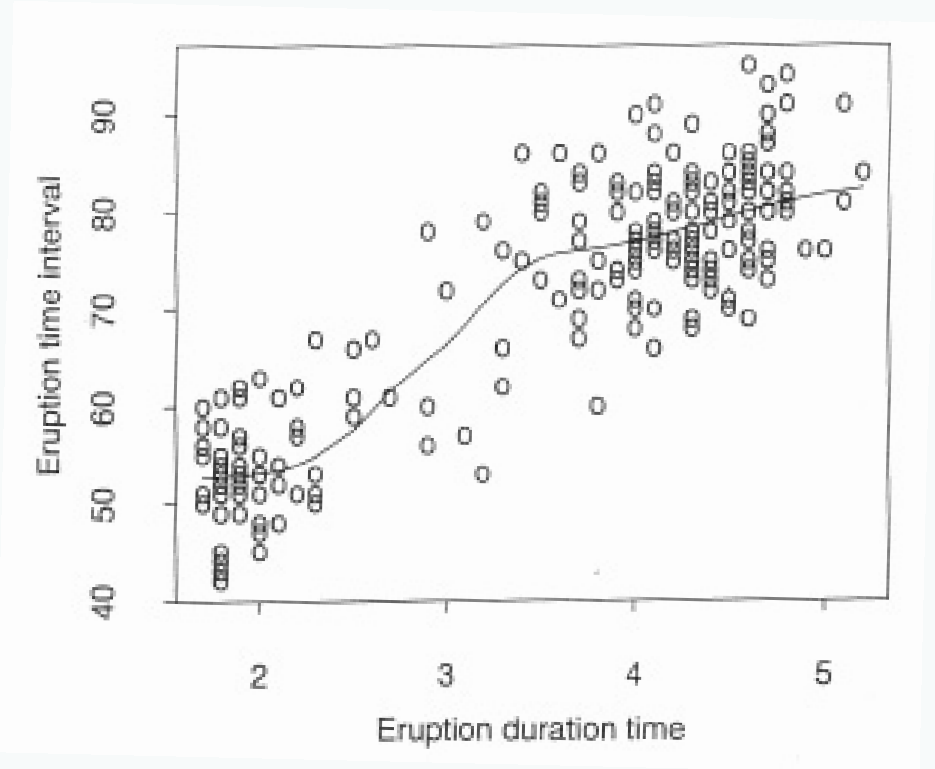
Interpretación del estimador de Nadaraya-Watson



En negro: modelo poblacional de datos simulados. **En gris:** estimador de Nadaraya - Watson. **El punto** representa la estimación puntual en $x = x_0$. **El área achurada** indica los pesos dados por un kernel Epanechnikov a cada observación.

MIDaS

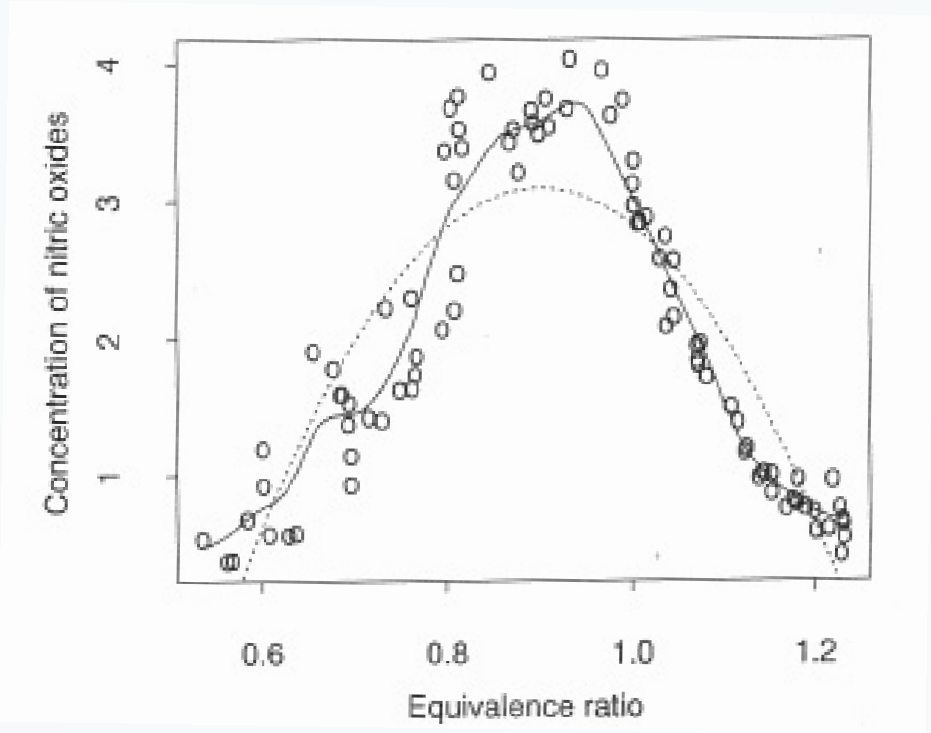
Algunos ejemplos:



Duración de erupciones del volcán Old Faithful Geyser. Estimador de Nadaraya-Watson con kernel Gaussiano y $h = 0,25$.

MiDaS

Nadaraya-Watson versus una función cuadrática



Niveles de óxido nítrico versus razón de equivalencia. **Linea sólida:** Nadaraya - Watson. **Línea punteada:** mínimos cuadrados para un polinomio de segundo orden.

MiDaS

Suavizamiento a través de splines

Compromiso entre la *fidelidad* y la *suavidad* de la curva:

Se busca una curva que minimice la función objetivo:

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - r(x_i))^2}_{\textit{fidelidad}} + \alpha \underbrace{\int_{x_{(1)}}^{x_{(n)}} [r''(u)]^2 \, du}_{\textit{suavidad}}.$$

El parámetro α determina qué tanto importa la suavidad por sobre la fidelidad.

Se demuestra que la solución a este problema de minimización corresponde a un *spline cúbico*.

MiDaS

Splines cúbicos:

Un spline cúbico corresponde a una función polinomial cúbica por trozos, en subintervalos definidos por nodos adyacentes $[x_{(i-1)}, x_{(i)}]$.

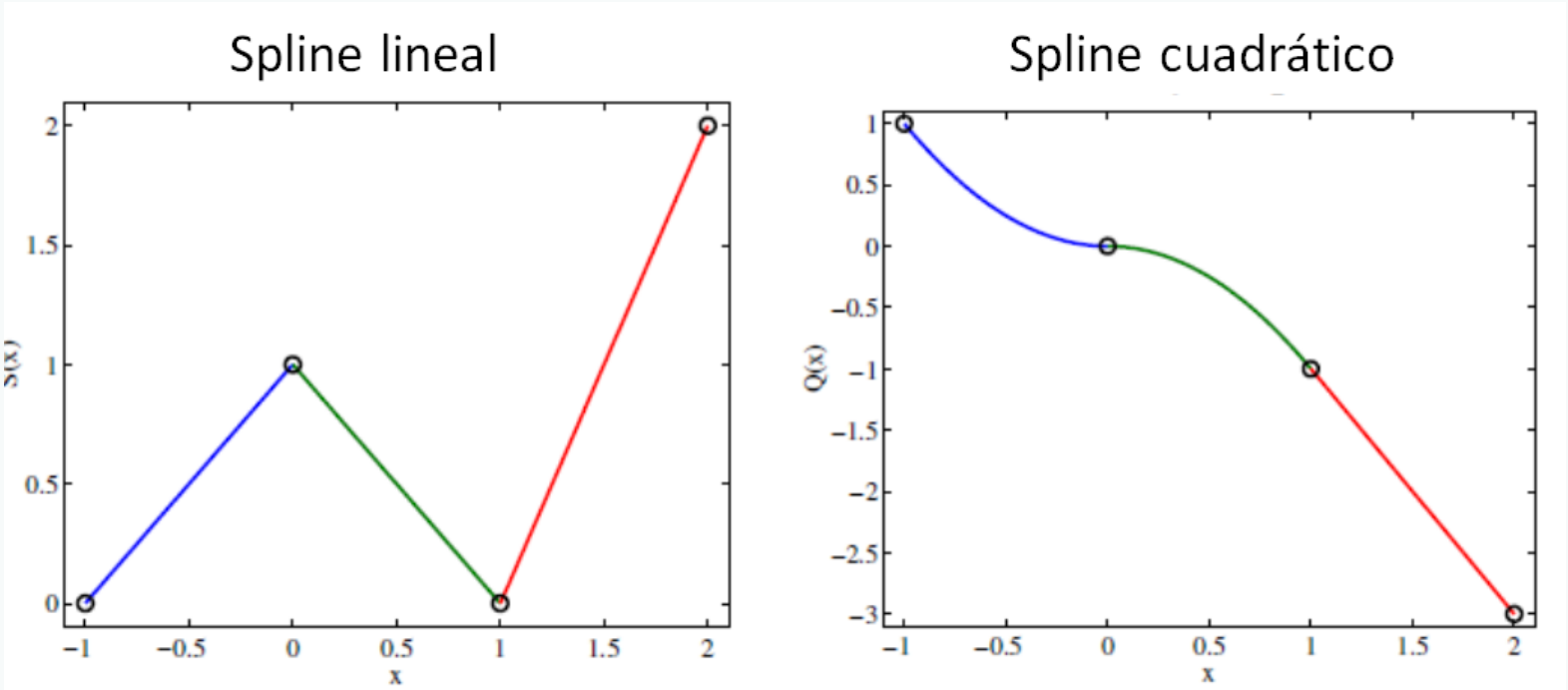
Posee las dos primeras derivadas continuas y la tercera corresponde a una función escalonada con saltos en los nodos. Esto asegura la suavidad de la curva.

Casos extremos:

- $\alpha = 0$: spline interpola, pasando por cada punto de la muestra.
- $\alpha = \infty$, spline se acerca a una linea recta.

MiDaS

Splines lineales y cuadráticos:



No son suficientemente suaves

MiDaS

Selección del parámetro de suavizamiento:

El valor de α puede ser escogido según el criterio de validación cruzada. Este criterio minimiza la suma de los cuadrados de las distancias entre la curva estimada y las observaciones.

En el Laboratorio de hoy tendrán la oportunidad de poner todo esto en práctica!

MiDaS