# Densidad y regresion no parametrica

Walter Sosa-Escudero

Universisad de San Andres y CONICET

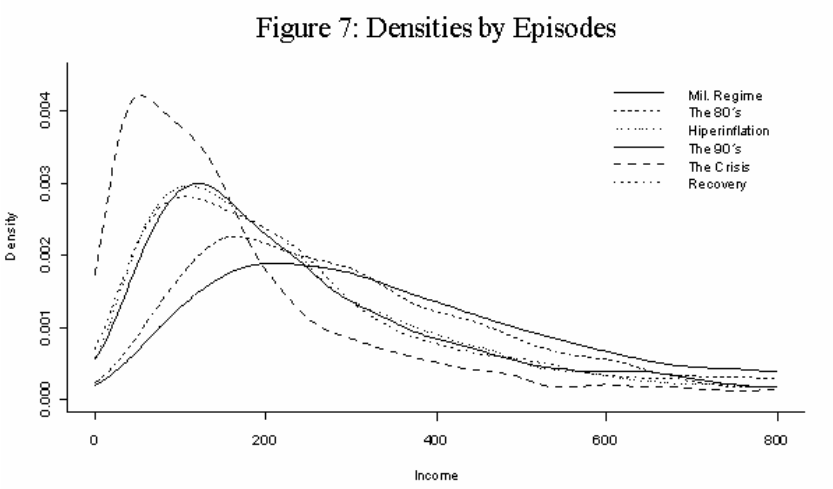


Figure 7: Densities by Episodes
Density
0.000
0.001
0.002
0.003
0.004
0
200
400
600
800
Income
Mil. Regime
The 80´s
Hiperinflation
The 90´s
The Crisis
Recovery

# Introduccion

$Y$, variable aleatoria continua.

$$F(y) = Pr(Y \leq y); \quad f(y) = \frac{dF(y)}{dy}$$

- Problema: estimar $f(y)$. Estimar $f(y_0)$.
- Porque? Distribucion del ingreso, Calificaciones de cursos, Rendimiento de activos.

# La aproximacion parametrica

$y \sim f(y;\theta)$, $y_i \sim f(y;\theta_0)$ i.i.d.

$f$ pertenece a una familia de distribuciones 'parametrizada' por $\theta$. No conocemos cuanto vale $\theta_0$.

- Si tuviesemos una estimacion $\hat{\theta}$, nuestra estimacion seria $\hat{f}(y_0;\hat{\theta})$
- El problema es *parametrico*: estimar $f(y_0)$ se traduce en un problema de estimar $\theta_0$ .
- Ventajas: confiables y potencialmente eficientes (ej, MV).
- Problema: requieren conocer $f()$. En muchas situaciones es precisamente lo que uno querria saber (ej, distribucion del ingreso).

# La aproximacion no-parametrica

Problema: estimar $f(y)$ a partir de $y_i \sim f(y)$ sin partir de supuestos sobre la forma de $f(y)$.

Aproximacion: metodo para estimar $f(y_0)$, que funciona para cualquier punto $y_0$. Replicable para cualquier otro punto el soporte.

- $Pr(Y = y_0) = 0$.
- Idea: $f(y_0)$ cuan 'frecuentes' son las realizaciones de $Y$ muy cerca de $y_0$.
- Dada una muestra iid $Y_i$, $i = 1, 2, \ldots, n$ que $f(y_0) > f(y_1)$ significa que deberia haber relativamente mas puntos cerca de $y_0$ que de $y_1$.

# El metodo de kernels

Propondremos estimadores que tienen la siguiente forma:

$$\hat{f}(y_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \, K\left( \frac{Y_i - y_0}{h} \right)$$

en donde $K(z)$ satisface:

1. $K(z) \geq 0$
2. $K(z) = K(-z)$ (simetrico en cero)
3. $\int K(s) \, ds = 1$ (integra a uno)
4. $\int sK(s) \, ds = 0$
5. $\int s^2 K(s) \, ds = \mu_2 < \infty$

$K(z)$ es un *kernel* y $h > 0$ es el *ancho de banda*.

# Intuicion

Kernel *gaussiano:*

$$K(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}$$

$K(z)$ satisface todos los criterios anteriores. Nuestro estimador es $\hat{f}(y_0) = 1/n \sum_{i=1} m_i$, con:

$$m_i \equiv \frac{1}{h}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Y_i - y_0}{h}\right)^2}$$

Para $y_0$ y $h$ dados, $m_i$ es decreciente en $|Y_i - y_0|$: $m_i$ mide cuan *cerca* esta $Y_i$ de $y_0$: cuanto mas pequeño es $|Y_i - y_0|$, mas grande es $m_i$.

Entonces:

$$\hat{f}(y_0) = \frac{1}{n}\sum_{i=1}^{n} m_i = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{Y_i - y_0}{h}\right)$$

mide *en promedio* cuan cerca estan todos los puntos de la muestra con respecto a $y_0$.

Mecanicamente:definir una *grilla* de valores del soporte, y repetir el procedimiento anterior para cada uno de los puntos del soporte:

# El rol de $h$

$$m_i \equiv \frac{1}{h}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{Y_i - y_0}{h}\right)^2}$$

A menor $h$, $Y_i$ es considerado como relativamente mas lejano de $y_0$, y viceversa. Para $h$ muy pequeño, todos los puntos son 'lejanos' y para $h$ exageradamente grande, todos los puntos son 'cercanos' con respecto a $y_0$.

# Kernel rectangular e histogramas

Kernel rectangular: $K(z) = \frac{1}{2}1[|z| < 0]$.

$\hat{f}(y_0) = 1/n \sum_{i=1} m_i$, con:

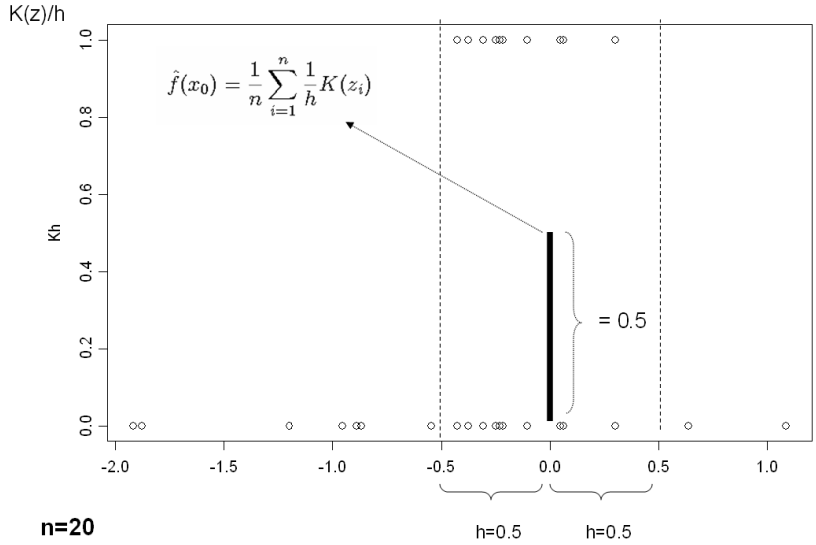
$$m_i \equiv \frac{1}{h}\frac{1}{2}1\left[\left|\frac{Y_i - y_o}{h}\right| < 1\right] = \frac{1}{2h}1\big[Y_i \in (Y_i - h, Y_i - h)\big]$$

Entonces:

$$\hat{f}(y_0) = \frac{1}{n}\sum_{i}^{n} m_i = \frac{\text{Proporcion de obs en } (y_0 - h, y_0 - h)}{2h}$$

Considera 'cerca' a todas las observaciones que no distan de $y_0$ en mas de $h$.

# Estimacion no-parametrica de densidades: metodo de kernels



# Propiedades del estimador: sesgo

$$
\begin{aligned}
E\left(\hat{f}(y_0)\right) &= E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{Y_i - y_0}{h}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}E\left[K\left(\frac{Y - y_0}{h}\right)\right] \qquad \text{(porque?)} \\
&= E\left[\frac{1}{h}K\left(\frac{Y - y_0}{h}\right)\right] \\
&= \int \frac{1}{h}K\left(\frac{s - y_0}{h}\right) f(s)\, ds
\end{aligned}
$$

$$
\begin{aligned}
E(\hat{f}(y_0)) &= \int \frac{1}{h} K\left(\frac{s-y_0}{h}\right) f(s)\, ds \\
&\quad \text{cambio de variables } s = y_0 + ht, \text{ con } ds = h\, dt \\
&= \int \frac{1}{h} K\left(\frac{y_0 + ht - y_0}{h}\right) f(y_0 + ht) h\, dt \\
&= \int K(t) f(y_0 + ht)\, dt \\
&\quad \text{sumamos y restamos } f(y_0) \\
&= f(y_0) + \int K(t) f(y_0 + ht)\, dt - \int f(y_0) K(t)\, dt \\
&= f(y_0) + \int \Big[ f(y_0 + ht) - f(y_0) \Big] K(t)\, dt
\end{aligned}
$$

De modo que para $h$ fijo el estimador es *sesgado*

# Propiedades del estimador: varianza

$$V(\hat{f}(y_0)) = V\left(\frac{1}{n}\sum_{i}^{n} m_i\right) = \frac{1}{n}V(m_i) \qquad \text{(porque?)}$$

$$\begin{aligned}
\frac{1}{n}V(m_i) &= \frac{1}{n}E(m_i^2) - \frac{1}{n}E(m_i)^2 \\
&= \frac{1}{n}\int \frac{1}{h^2}K^2\left(\frac{s-y_0}{h}\right) f(s)\, ds - \frac{1}{n}\left[\int K(t) f(y_0+ht) dt\right]^2 \\
&= \frac{1}{nh}\int K^2(t) f(y_0+ht)\, dt - \frac{1}{n}\left[\int K(t) f(y_0+ht)\, dt\right]^2
\end{aligned}$$

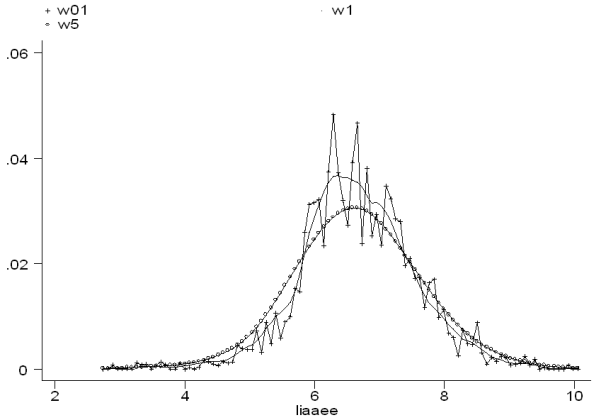
# Trade-off sesgo y varianza

Entonces

$$
\begin{aligned}
E\Big(\hat{f}(y_0)\Big) &= f(y_0) + \int \Big[f(y_0+ht) - f(y_0)\Big] K(t)\, dt \\
V\Big(\hat{f}(y_0)\Big) &= \frac{1}{nh}\int K^2(t) f(y_0+ht)\, dt - \frac{1}{n}\left[\int K(t) f(y_0+ht)\, dt\right]^2
\end{aligned}
$$

Trade off: $h \to 0$ implica $E\Big(\hat{f}(y_0)\Big) \to 0$ y $V\Big(\hat{f}(y_0)\Big) \to \infty$

# Comparacion de anchos de banda



# Propiedades del estimador: consistencia

*Resultado:* si cuando $n \to \infty$, la varianza y el sesgo de un estimador tienden a cero, entonces el estimador es consistente (convergencia en media cuadratica). Son condiciones *suficientes* para la consistencia.

- Supongamos que $h = h_n$
- Si $h_n \to 0$ y $nh_n \to \infty$ entonces el sesgo y la varianza tienden a cero, de modo que el estimador es consistente.
- La consistencia requiere que $h$ vaya cero a una tasa menor a la que $n$ va a infinito.
- Notar que la propiedad de consistencia *no* depende del kernel elegido!

# Ancho de banda optimo

$$ECM\ (\hat{f}(y_0)) = E(\hat{f}(y_0) - f(y_0))^2 = \text{Sesgo}^2(\hat{f}) + V(\hat{f})$$

Problema: queremos encontrar un $h$ optimo que funcione para todo el soporte, no solo para un punto:

$$ECMI\ (\hat{f}(y_0)) = \int \text{Sesgo}^2(\hat{f}) + V(\hat{f})dt$$

Idea: minimizar ECMI con respecto a $h$. Problema: depende de conocer $f()$!!

Se puede mostrar que $h$ optimo se puede aproximar razonablemente como:

$$h = cn^{-1/5}$$

con

$$c = \frac{\mu_2^2 \int (f^{(2)}(s))^2 dx}{\int K^2(t) dt}$$

en donde se ve explicitamente que el $h$ optimo depende de $f()$, a traves de sus segundas derivadas y del kernel utilizado.

Si la verdadera distribucion $f(y)$ fuese normal con media $\mu$ y varianza $\sigma^2$ y el kernel utilizado fuese gaussiano:

$$h^* = 1.06\sigma n^{-1/5}$$

Silverman (1996): $h^*$ funciona correctamente aun cuando $f(y)$ no es normal, siempre y cuando la densidad verdadera no sea bimodal o demasiado sesgada.

Hay miles de mejoras y aproximaciones. La de Shether y Jones (1991) es muy popular.

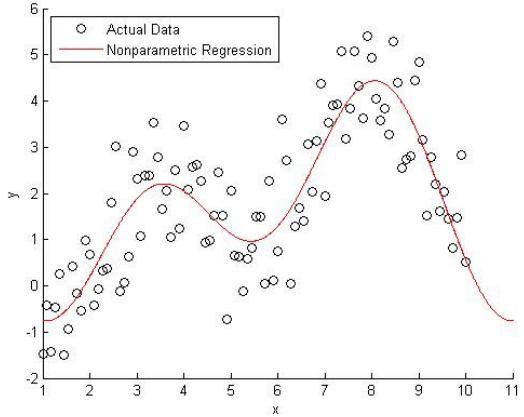
Inspeccion visual: de acuerdo a como funciona el trade off, empezar con un $h$ chico y aumentarlo progresivamente. Importancia de estimar varios.

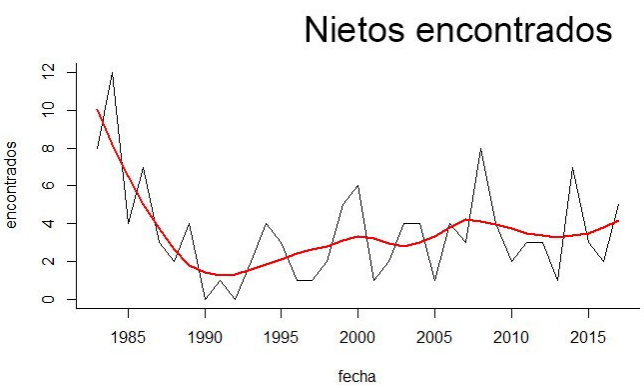
Machine learning: cross validation!

# Regresion no parametrica

$$E(Y|X = x_0) = g(x_0)$$

- Enfoque parametrico: $g(x_0) = g(x_0, \theta)$. Ej: $x_0\beta$.
- Enfoque no parametrico: estimar $g(x_0)$ directamente


Actual Data
Nonparametric Regression
y
x
6
5
4
3
2
1
0
-1
-2
1
2
3
4
5
6
7
8
9
10
11


Nietos encontrados
encontrados
0
2
4
6
8
10
12
1985
1990
1995
2000
2005
2010
2015
fecha

# Estimador de Nadaraya/Watson

Para una muestra $(X_i, Y_i), i = 1, \ldots, n$, iid.

$$\hat{g}(x_0) = \sum_{i=1}^{n} w_i Y_i$$

con

$$w_i \equiv \frac{K\left(\frac{X_i - x_0}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)}$$

Intuicion: promedio local ponderado por la distancia a $x_0$

# $k-$ vecinos cercanos

$N_k(x_0)$ : conjunto de $k$ las observaciones mas cercanas a $x_0$.

$$\hat{g}_{\mathsf{knn}}(x_0) = \frac{1}{k} \sum_{i=1}^{n} 1[x_i \in N_k(x_0] \; Y_i$$

Intucion: promedio de las $k$ $Y_i$'s con $X_i$ mas cerca de $x_0$

# Regresion local

$$\hat{g}_{|}(x_0) = \hat{\alpha} + \hat{\beta} x_o$$

en donde $\hat{\alpha}$ y $\hat{\beta}$ minimizan

$$\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)\left(Y_i - \hat{\alpha} + \hat{\beta}(X_i - x_0)\right)^2$$

Intuicion: regresion local

# La maldicion de la dimensionalidad

$f(y, x)$ ?

- Caso univariado. Kernel? Cerca de $x_0$
- Multivariado. Cerca de $(x_0, y_o)$.
- Punto $f(x)$ integra a 1, $f(x, y)$.... tambien.
- Problema: para una nocion dada de 'cerca', la probabilidad de estar cerca cae dramaticamente con la cantidad de dimensiones.
- Ejemplo: $p$ variables independientes $N(0, 1)$. Cerca de 0: a menos de 0.5. $p = 1, 0.3833, p = 2, 0.118, p = 3, 0.031$
- Maldicion de la dimensionalidad: la cantidad de informacion 'cerca' cae rapidamente con la dimensionalidad.

Muy dificil extrapolar metodos no parametricos para dimensionalidad alta.