Estadística no paramétrica

Regresión no paramétrica I: Métodos basados en kernel

Jorge de la Vega Góngora

Departamento de Estadística,
Instituto Tecnológico Autónomo de México

14 de abril de 2023

Introducción

Introducción I

- En esta sección estudiaremos el suavizamiento de gráficas de dispersión, como parte de los conceptos relevantes de regresión no paramétrica. Nos concentraremos en el caso bivariado, aunque la notación será general y se extiende fácilmente a más dimensiones.
- Los modelos de regresión no paramétrica se han vuelto comunes dadas las facilidades computacionales actuales. Una de las grandes ventajas de los modelos no paramétricos es que permite estimar modelos no lineales fácilmente. El modelo de estimación a través de kerneles de densidad puede extenderse a la estimación de curvas.
- El principio de modelado es fácil de plantear: Si $(X_1, Y_1), \ldots, (X_n, Y_n)$ es un conjunto de n pares independientes de puntos de un vector bivariado (X, Y), definimos la *función de regresión* como la media condicional

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

Introducción II

- Sea

$$Y_i = m(X_i) + \epsilon \quad \text{para} \quad i = 1, \ldots, n$$

donde ϵ_i son errores con media E$(\epsilon_i) = 0$ y Var$(\epsilon_i) = \sigma^2$ constante. El modelo clásico de regresión lineal, que es paramétrico, supone que $Y_i|\mathbf{x} \sim \mathcal{N}\left(m(\mathbf{x}), \sigma^2\right)$ o equivalentemente, que los errores son normales con media 0 y varianza σ^2 y adicionalmente, que la función de regresión es una función lineal, $m(\mathbf{x}) = \beta'\mathbf{x}$.

- En un modelo de regresión no paramétrica, se relaja el supuesto de linealidad, suponiendo que la función m es una función *suave*. El costo es computacional y en algunos casos, un resultado más difícil de entender, pero la ganancia es una estimación más precisa de la función de regresión.
- A veces parece que los modelos de regresión no paramétrica, y en general las técnicas de suavizamiento, no tienen mucho sustento teórico, pero en realidad estamos hablando de un poco de análisis funcional, y ajustando funciones. Aquí tomaremos un sentido práctico tratando de fundamentar las ideas principales, pero no entraremos mucho en detalles teóricos.

Ejemplo: Análisis de prestigio ocupacional (Blishen, 1976) I

- Como ejemplo ilustrativo de las ideas del ajuste de suavizamiento y regresión no paramétrica, consideramos el siguiente: se obtienen puntajes sobre el prestigio de 102 ocupaciones en Canadá a través de una encuesta, y se ajusta un modelo de regresión lineal múltiple a los puntajes asignados (variable `Prestige`). Además, se relacionan dichos puntajes al ingreso promedio (`Income`) y al nivel de educación promedio (`Education`) de esa ocupación. Los datos se encuentran en el archivo `Prestige.txt`

```
datos <- read.delim("https://raw.githubusercontent.com/jvega68/ENP/main/datos/Prestige.txt", sep = "", header = T)
head(datos)
```

```
                    education income women prestige census type
GOV.ADMINISTRATORS      13.11  12351 11.16     68.8   1113 prof
GENERAL.MANAGERS        12.26  25879  4.02     69.1   1130 prof
ACCOUNTANTS             12.77   9271 15.70     63.4   1171 prof
PURCHASING.OFFICERS     11.42   8865  9.11     56.8   1175 prof
CHEMISTS                14.62   8403 11.68     73.5   2111 prof
PHYSICISTS              15.64  11030  5.13     77.6   2113 prof
```

```
lm(prestige ~ income + education, data = datos)
```

```

Call:
lm(formula = prestige ~ income + education, data = datos)

Coefficients:
(Intercept)       income    education
  -6.847779     0.001361     4.137444
```

Ejemplo: Análisis de prestigio ocupacional (Blishen, 1976) II

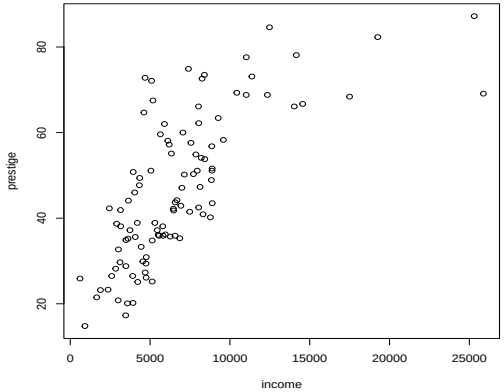
- El propósito de este análisis fue predecir puntajes para otras ocupaciones no incorporadas en las encuestas y de los que no se tenía valor.
- El modelo de regresión lineal ajusta un plano a los datos, con un modelo de la forma

$$E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

- Para poder visualizar el modelo, consideremos una relación más simple, veamos la relación entre `prestige` y `income`, en donde un modelo lineal no necesariamente es el más apropiado:

```
with(datos, plot(income, prestige))
```

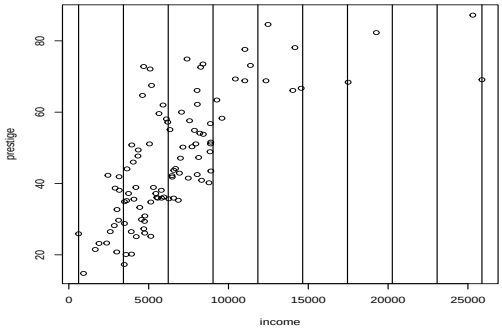
Ejemplo: Análisis de prestigio ocupacional (Blishen, 1976) III



- ¿Qué modelo paramétrico sugieren que se ajuste?

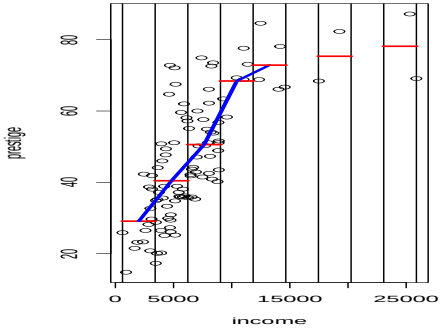
Ejemplo: Análisis de prestigio ocupacional (Blishen, 1976) IV

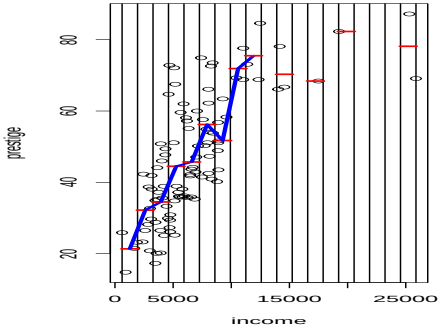
- Un procedimiento simple (pero que es la base para muchos de los métodos que siguen) para 'ajustar' una curva es similar al que usamos para construir un histograma, se conoce como *binning* o promedio local:
  1. Se particiona el dominio de X para generar bandas en el diagrama de dispersión de puntos:



Ejemplo: Análisis de prestigio ocupacional (Blishen, 1976) V

2. Cuando las bandas creadas son pequeñas, se puede obtener una aproximación a $\mathrm{E}(Y|X)$ en cada banda calculando $\bar{y}|(x$ en banda $i)$.
3. Finalmente unimos los puntos centrales de cada *bin*. Noten que algunos bins no tienen puntos, por lo que no se puede calcular el nivel en ese bin. Se puede entonces unir las medias de los bins que sí tienen datos.





Ejemplo: Análisis de prestigio ocupacional (Blishen, 1976) VI

- El procedimiento anterior, con variaciones en complejidad, y con consideraciones a casos de diferentes situaciones, es lo que da origen a una variedad de métodos de solución.
- Como en los casos de cualquier estimación, tenemos un intercambio entre sesgo y varianza, y el factor del que depende la ponderación de uno u otro está relacionado con el tamaño del ancho de banda considerado.

Ejemplos de modelos de regresión no paramétrica I

Algunos ejemplos de modelos que se considerarán son los siguientes:

- **Estimadores basados en kernel**: como ejemplos, veremos Nadaraya-Watson, Gasser-Müller y polinomios locales.
- **Modelos de vecinos más cercanos (*nearest neighbors*)**: este incluye `loess`.
- **splines**
- **Modelos aditivos**: Son modelos de la forma $y = \alpha + \sum_{i=1}^{p} f_i(x_i) + \epsilon$.
- **onduletas** o *wavelets*

Estimadores basados en kernel

Estimadores basados en kernel I

- La estimación basada en kernel usa la idea de colocar un kernel de densidad K en cada uno de los puntos X_i de la muestra para ponderar los puntos Y_i que están en una vecindad de ese valor.
- La idea esencial es que al estimar $\hat{y}_0 = \hat{m}(x_0)$, es deseable dar mayor peso a las observaciones que están cerca de un punto focal x_0 y menos peso a los que estén lejos.
- Los pesos que se considerarán son de la forma $w_i = K\left(\frac{X_i - x_0}{h}\right)$, y entonces se procede a calcular el estimado como

$$\hat{y}_0 = \hat{m}(x_0) = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

- Si recordamos, en regresión lineal simple, el estimador de mínimos cuadrados es *BLUE* (o *MELI*: mejor estimador lineal insesgado) y tiene la característica de ser una combinación lineal de las observaciones, es decir, tiene la forma:

$$\hat{m}(x) = \sum_{i=1}^{n} a_i Y_i$$

con ciertos pesos a_i. Entonces el estimador de kernel nos da también una combinación lineal de las observaciones, y los pesos dependen de las observaciones X_i.

Funciones kernel comunes en suavizamiento I

- En el caso de suavizamiento, a diferencia de la estimación de funciones de densidad, la función kernel K *no necesariamente* tiene que cumplir las propiedades de densidad. Por ejemplo, se puede permitir que tome valores negativos, y en algunos casos esto es necesario para alcanzar estimadores óptimos en un sentido asintótico.
- Un ejemplo de kernel que se usa en el contexto de suavizamiento, es el kernel beta:

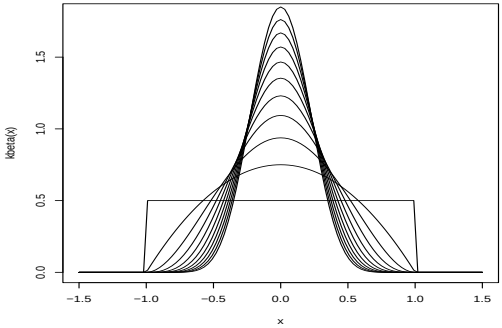
$$K(x) = \frac{1}{B(1/2, \gamma + 1)}(1 - x^2)^\gamma I(|x| \leq 1), \quad \gamma = 0, 1, 2, ...$$

El parámetro γ hace flexible este kernel:

- $\gamma = 0$ corresponde a una uniforme
- $\gamma = 1$ corresponde al kernel de Epanechnikov
- $\gamma = 2$ es el kernel bipeso
- $\gamma = 3$ es el kernel tripeso
- Cuando γ es suficientemente grande, se converge al kernel gaussiano.

Funciones kernel comunes en suavizamiento II

```
kbeta <- function(x,gama = 0){
        1/beta(0.5,gama+1) * (1-x^2)^(gama)*ifelse((x <= 1) & (x >= -1),1,0)
}
curve(kbeta(x),from =-1.5, to = 1.5, ylim = c(0,1.8))
for(i in 1:10)curve(kbeta(x, gama = i),from = -1.5, to = 1.5, add = T)
```

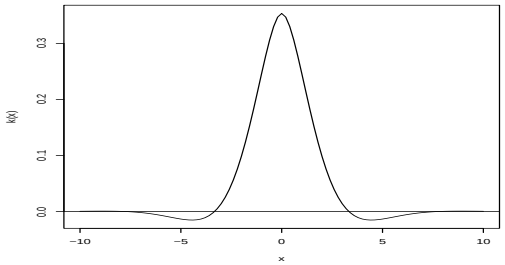


Funciones kernel comunes en suavizamiento III

- Un ejemplo de un kernel generalizado (que puede tomar valores negativos) es la siguiente función:

$$K(x) = \frac{1}{2}\exp(-|x|/\sqrt{2})\sin(|x|/\sqrt{2} + \pi/4)$$

```
k <- function(x){
        0.5*exp(-abs(x)/sqrt(2))*sin(abs(x)/sqrt(2) + pi/4)
}
curve(k(x),from =-10, to = 10)
abline(h=0)
```



Estimadores de Nadaraya-Watson (1964) I

- El estimador de Nadaraya-Watson de $m(x)$ se define como:

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_h(X_i - x) Y_i}{\sum_{i=1}^{n} K_h(X_i - x)}$$

con $K_h(u) = K(u/h)$ es una función kernel.

Teorema (Estimador de Nadaraya-Watson (1964))

Para una x fija, el valor $\hat{\theta}$ que minimiza la función

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

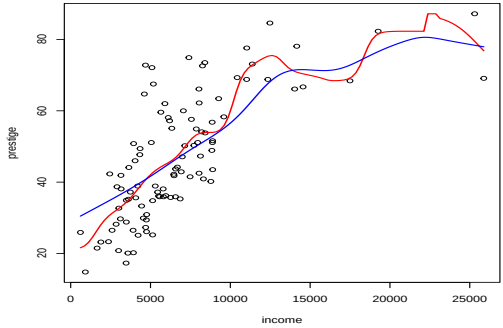
es una combinación lineal de las observaciones Y_i: $\hat{\theta} = \sum_{i=1}^{n} a_i Y_i$. Además, los ponderadores óptimos están dados por $a_i = \frac{K_h(X_i - x)}{\sum_{i=1}^{n} K_h(X_i - x)}$

Estimadores de Nadaraya-Watson (1964) II

- Noten que la función a optimizar es un error cuadrático medio ponderado, así que la forma de demostrar el resultado es similar a obtener el estimador de mínimos cuadrados ponderados en regresión lineal.

Ejemplo 1

```
plot(income, prestige)
lines(ksmooth(income, prestige, "normal", bandwidth = 2000), lwd = 2, col = "red")
lines(ksmooth(income, prestige, "normal", bandwidth = 5000), lwd = 2, col = "blue")
```



Ejemplo 2

```
x <- sort(runif(200))
y <- sin(4*pi*x) + 0.3*rnorm(200)
plot(x,y)
lines(ksmooth(x, y, "normal", bandwidth = 0.07), lwd = 3, col = "red")
lines(ksmooth(x, y, "normal", bandwidth = 0.14), lwd = 3, col = "green")
lines(ksmooth(x, y, "normal", bandwidth = 0.21), lwd = 3, col = "blue")
```