Trabajo Practico Final - Barragan, Horn, Laria

2024-07-11

Teórico

a)

$$\begin{pmatrix} \hat{m}_h(X_1) \\ \vdots \\ \hat{m}_h(X_n) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n Y_i \omega_{i,h}(X_1) \\ \vdots \\ \sum_{i=1}^n Y_i \omega_{i,h}(X_n) \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} \omega_{1,h}(X_1) & \cdots & \omega_{n,h}(X_1) \\ \vdots & \ddots & \vdots \\ \omega_{1,h}(X_n) & \cdots & \omega_{n,h}(X_n) \end{pmatrix}}_{\text{S}} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

$$\text{donde } S_{ij} = \omega_{j,h}(X_i)$$

b)

$$\frac{\hat{m}_h(X_i) - Y_i \omega_{i,h}(X_i)}{1 - \omega_{i,h}(X_i)} = \frac{\left(\sum_{j=1}^n Y_j \omega_{j,h}(X_i)\right) - Y_i \omega_{i,h}(X_i)}{1 - \omega_{i,h}(X_i)}$$

$$= \frac{\left(\sum_{j=1}^n Y_j \frac{K\left(\frac{X_j - X_i}{h}\right)}{\sum_{l=1}^n K\left(\frac{X_l - X_i}{h}\right)}\right) - Y_i \frac{K\left(\frac{X_i - X_i}{h}\right)}{\sum_{l=1}^n K\left(\frac{X_l - X_i}{h}\right)}}{1 - \frac{K\left(\frac{X_i - X_i}{h}\right)}{\sum_{l=1}^n K\left(\frac{X_l - X_i}{h}\right)}} = \frac{\left(\sum_{j=1}^n Y_j \frac{K\left(\frac{X_j - X_i}{h}\right)}{\sum_{l=1}^n K\left(\frac{X_l - X_i}{h}\right)}\right) - Y_i \frac{K\left(\frac{X_i - X_i}{h}\right)}{\sum_{l=1}^n K\left(\frac{X_l - X_i}{h}\right)}}{\frac{\sum_{l=1}^n K\left(\frac{X_l - X_i}{h}\right) - K\left(\frac{X_i - X_i}{h}\right)}{\sum_{l=1}^n K\left(\frac{X_l - X_i}{h}\right)}}$$

$$= \frac{\left(\sum_{j=1}^n Y_j K\left(\frac{X_i - X_j}{h}\right)\right) - Y_i K\left(\frac{X_i - X_i}{h}\right)}{\left(\sum_{\ell=1}^n K\left(\frac{X_\ell - X_i}{h}\right)\right) - K\left(\frac{X_i - X_i}{h}\right)} = \sum_{j=1, j\neq i}^n Y_j \frac{K\left(\frac{X_i - X_j}{h}\right)}{\sum_{\ell=1, \ell\neq i}^n K\left(\frac{X_\ell - X_i}{h}\right)} = \hat{m}_h^{-i}(X_i)$$

Por definición,

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

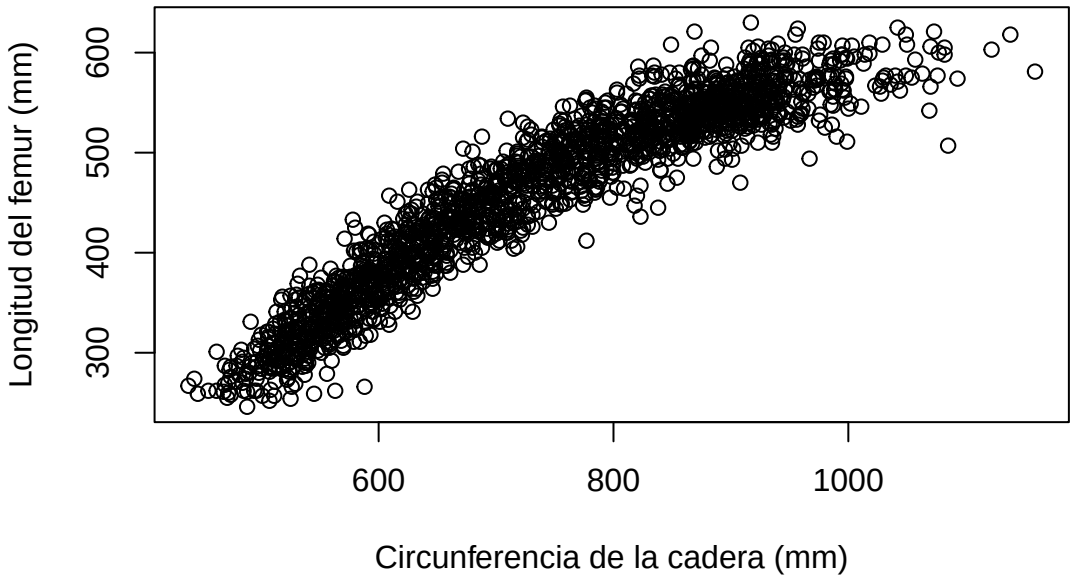
Enchufando lo probado anteriormente,

$$CV(h) = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \frac{\hat{m}_h(X_i) - Y_i \omega_{i,h}(X_i)}{1 - \omega_{i,h}(X_i)}\right)^2 = \frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i - Y_i \omega_{i,h}(X_i) + Y_i \omega_{i,h}(X_i) - \hat{m}_h(X_i)}{1 - \omega_{i,h}(X_i)}\right)^2$$

$$= \frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i - \hat{m}_h(X_i)}{1 - \omega_{i,h}(X_i)}\right)^2 = \frac{1}{n} \sum_{i=1}^n \frac{(Y_i - \hat{m}_h(X_i))^2}{(1 - \omega_{i,h}(X_i))^2}$$

Práctico

a)



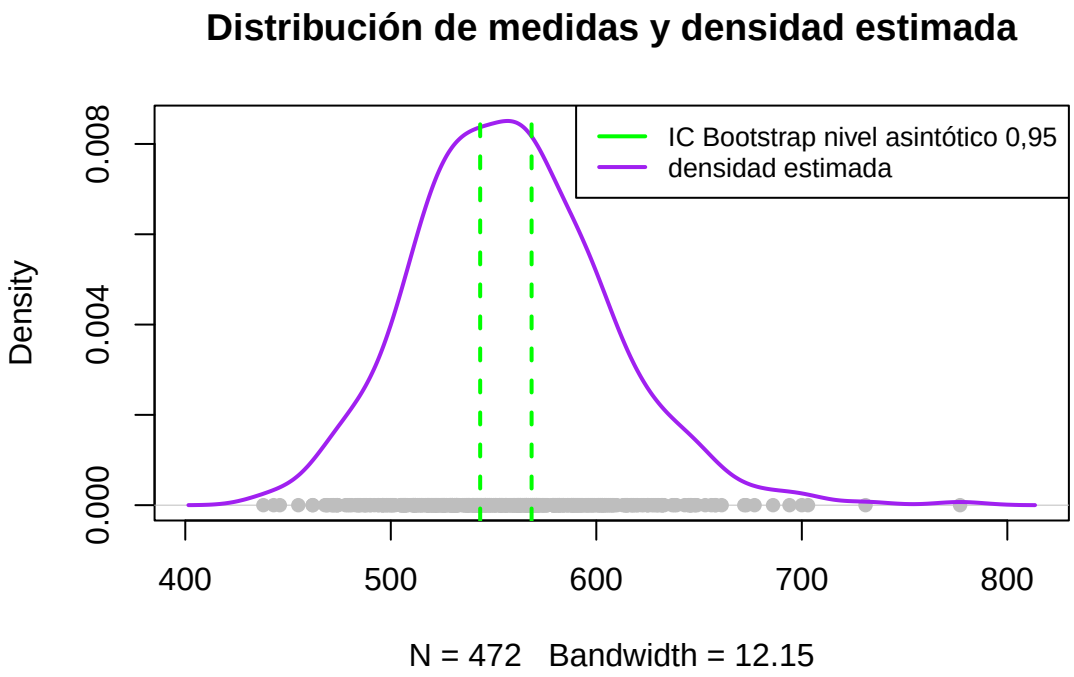
Vemos que este diagrama de dispersión nos indica que existe una relacion creciente entre las variables HIP.CIRCUMFERENCE y BUTTOCK.KNEE.LENGTH. Es lógico que esten correlacionadas positivamente estas variables ya que a mayor circunferencia de cadera seguramente se corresponda con una mayor longitud del fémur.

b)

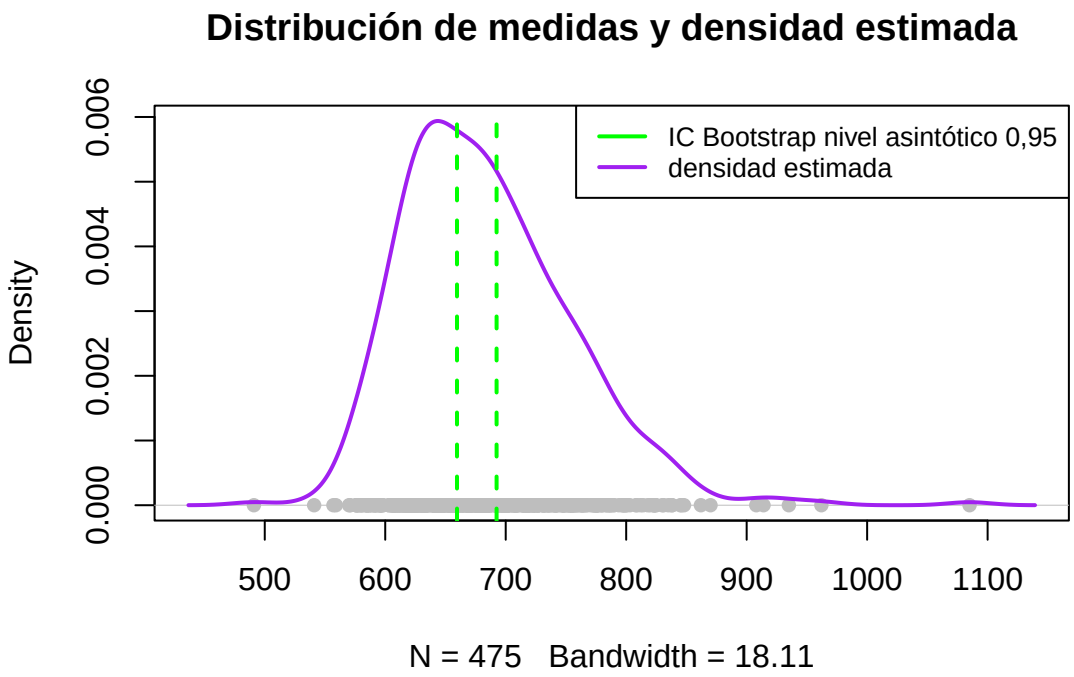
```
## [1] "Las estimaciones para cada grupo son: "

## La estimacion del primer grupo etario (0 y 81 meses) es: 556 mm

## La estimacion del segundo grupo etario ( 81 y 128 meses) es: 676 mm

## La estimacion del tercer grupo etario ( 128 y 168 meses) es: 799 mm

## La estimacion del cuarto grupo etario (mas de 168 meses) es: 904 mm
```

Intervalos de confianza

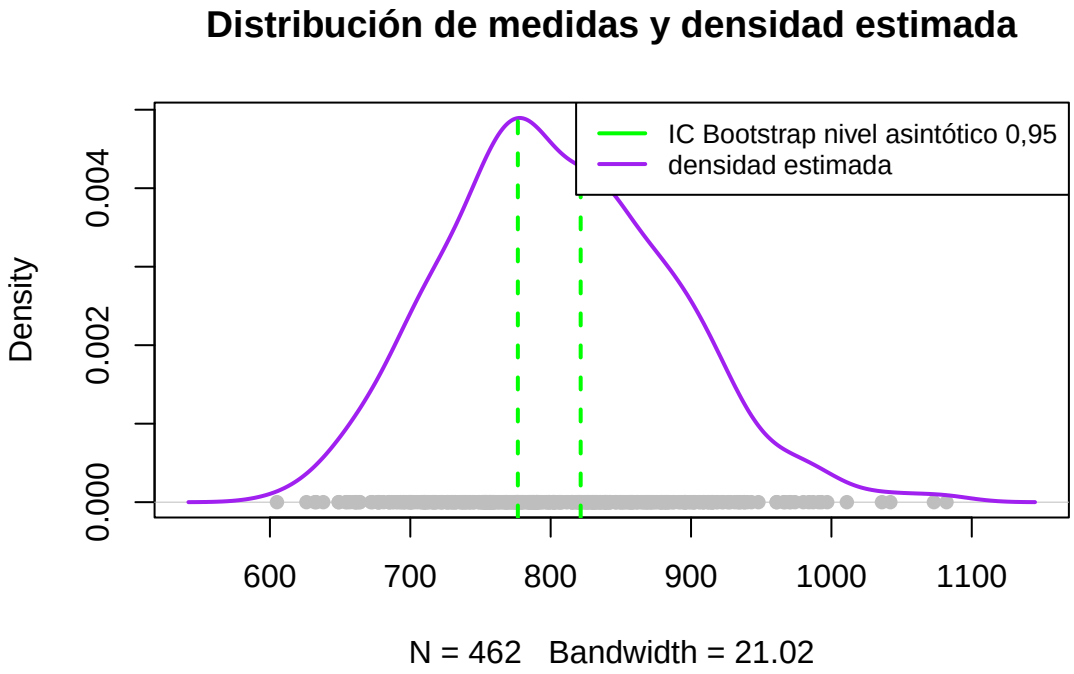
```
## 1 º grupo: ( 543.4926 , 568.5074 )
```



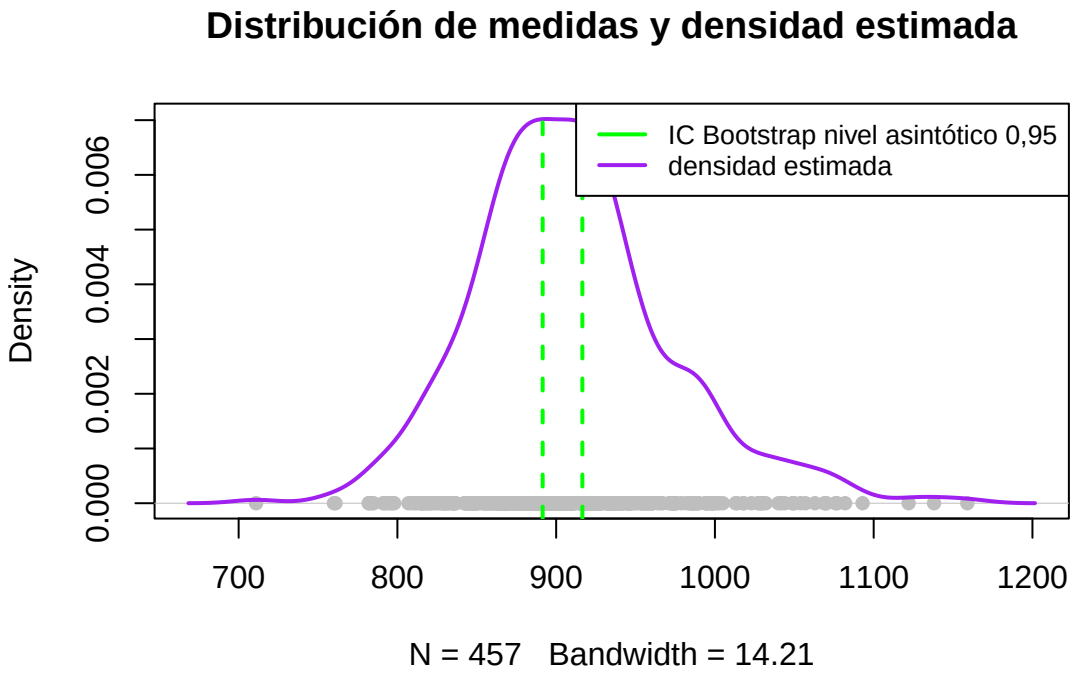

```
## 2 º grupo: ( 659.5754 , 692.4246 )
```




```
## 3 º grupo: ( 776.6165 , 821.3835 )
```




```
## 4 º grupo: ( 891.5004 , 916.4996 )
```



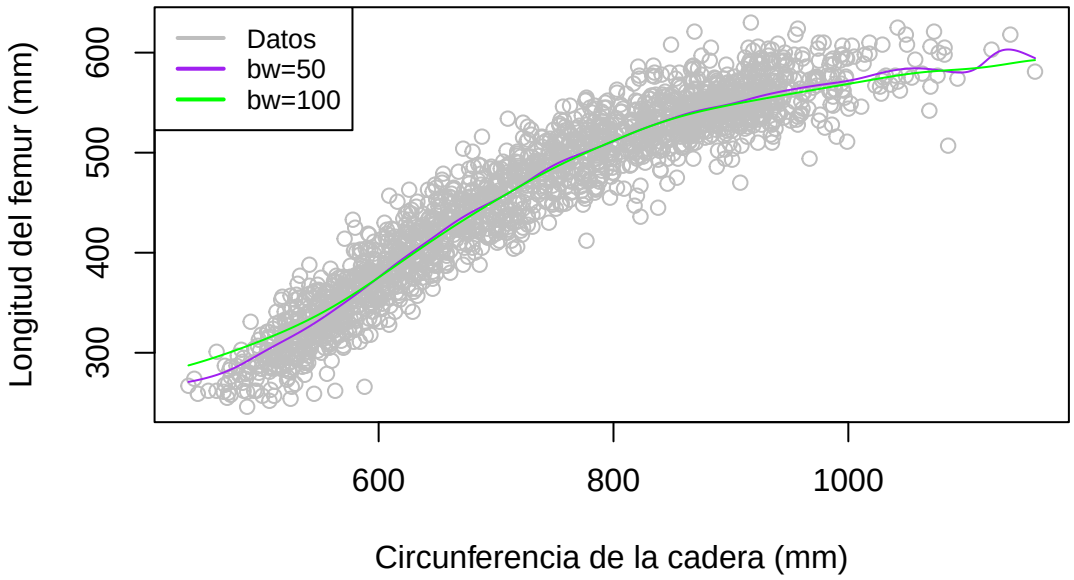
Algo que podemos ver a simple vista y sin calculos previos es que, a mayor dispersión de los datos, mayor longitud tiene su intervalo de confianza Bootstrap. A priori, esto no es tan directo porque los ICs son calculados con el desvío estándar de cada conjunto de estimadores Bootstrap. Una pregunta interesante podría ser si efectivamente esto es así.

Para calcular cada uno de los intervalos de confianza de nivel 0.95 (normales y aproximados) para la

mediana de cada grupo, primero estimamos la mediana de ese grupo, y luego estimamos el desvío estándar del estimador utilizado a partir de generar 5000 remuestras bootstrap.

c)

i)



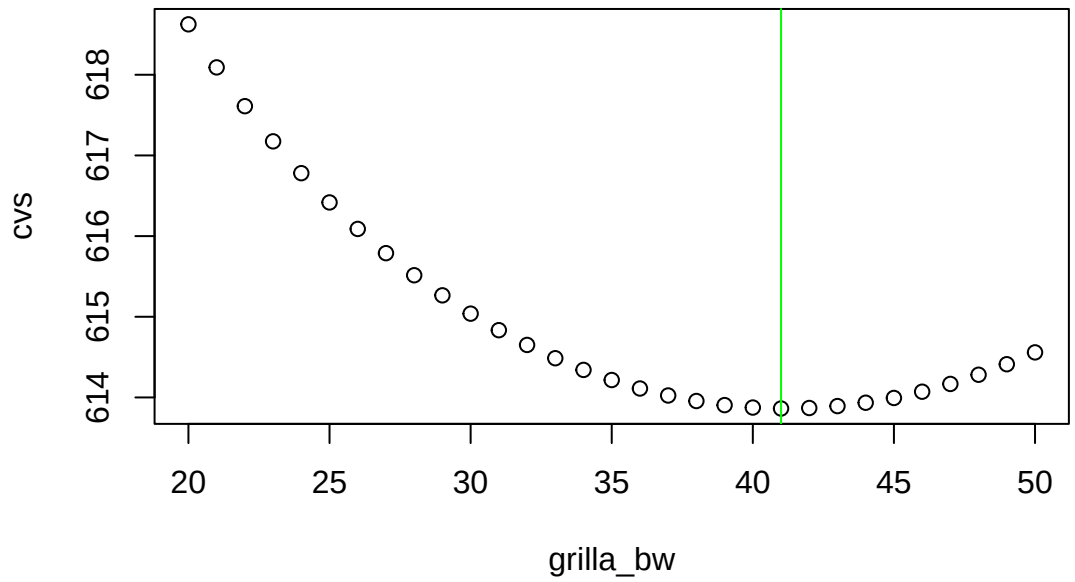
Calculemos el ECM correspondiente a cada ventana y quedemonos con la que minimice.

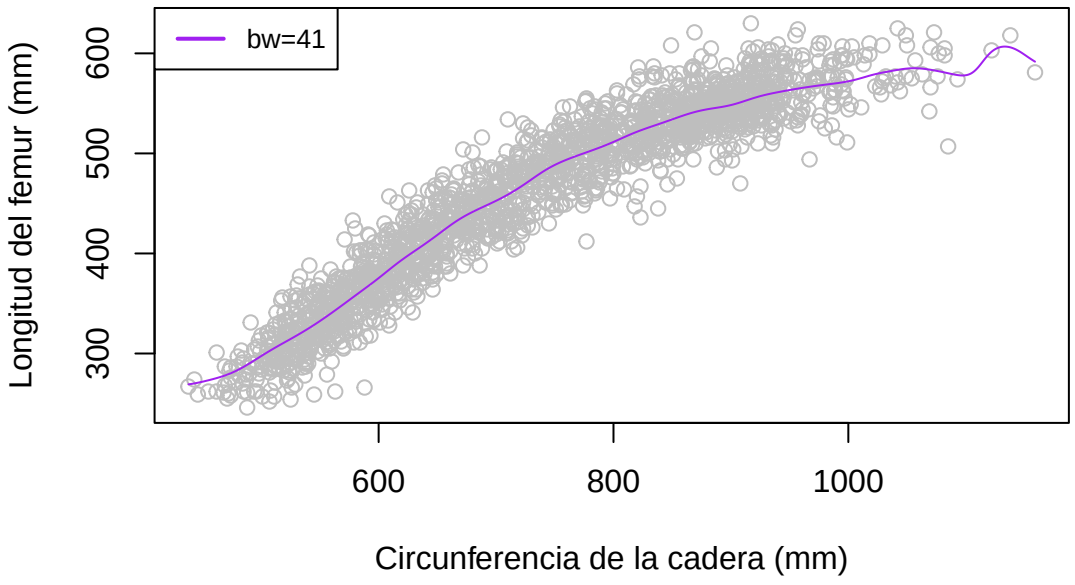
```
## CV(50) = 614.5587
```

```
## CV(100) = 643.833
```

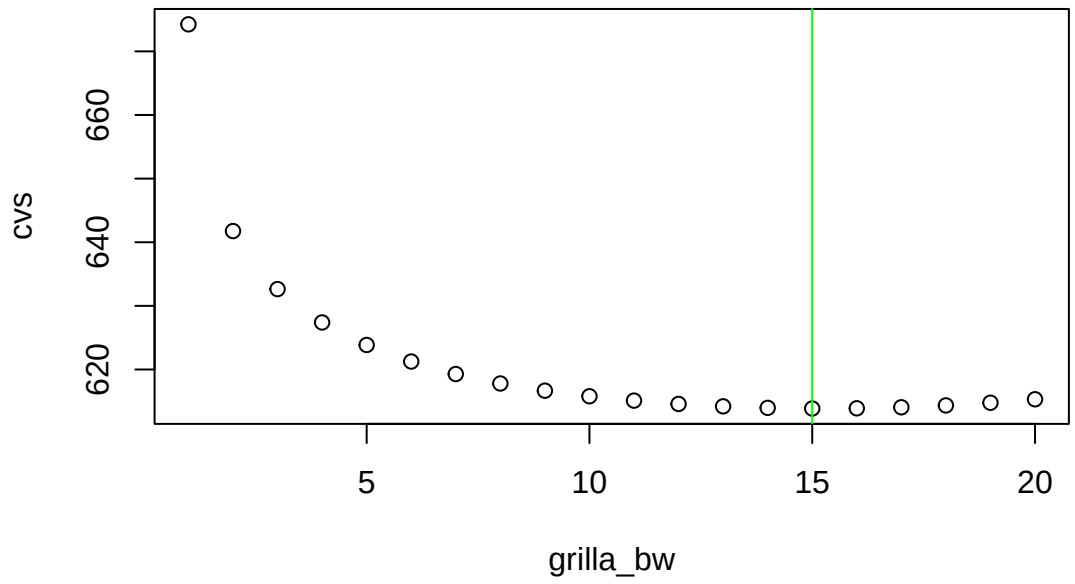
La ventana $bw = 50$ es la que minimiza el error de CV entre las dos opciones. Nos quedamos con $bw = 50$.

ii)

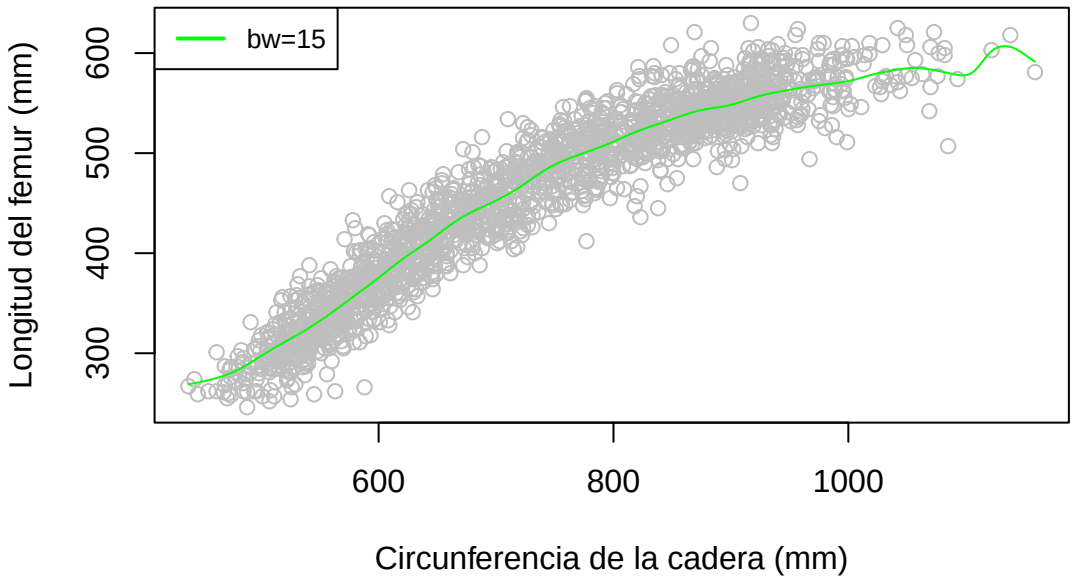

cvs
614
615
616
617
618
20
25
30
35
40
45
50
grilla_bw


bw=41
Longitud del femur (mm)
300
400
500
600
600
800
1000
Circunferencia de la cadera (mm)

iii)

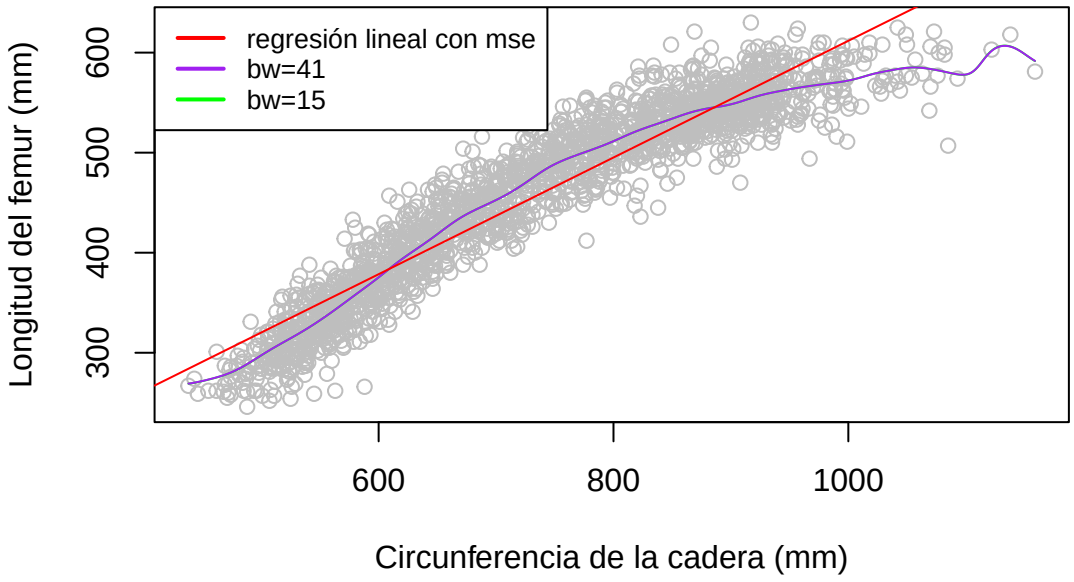


Cambiamos la grilla porque ksmooth "escala" los kernels, entonces al hacerlo a mano se nos corrre la bw óptima. Vemos que obtenemos la misma regresión para los dos algoritmos, con distintas ventanas óptimas.



iv)

Usamos las fórmulas de la pendiente y ordenada óptimas en el sentido de cuadrados mínimos porque la matriz de diseño es inversible (se ve fácil porque los valores de la covariable HIP CIRCUMFERENCE no son todos iguales, entonces con eso estamos).



Vemos que las estimaciones de Nadaraya-Watson se superponen para las ventanas $bw = 15; 41$. Esto es porque la implementación de Ksmooth escala los núcleos con el fin de que sus cuartiles se encuentran en $+ - 0.25bw$.

Si usamos la implementación de ksmooth, no cambiamos la grilla. Si en cambio usamos nuestra implementación, usamos la grilla que graficamos anteriormente porque ahora el mínimo de la función objetivo (CV(h)) se encuentra en otro lado.

A simple vista no elegiríamos la regresión lineal. Veamos los MSEs de cada regresión.

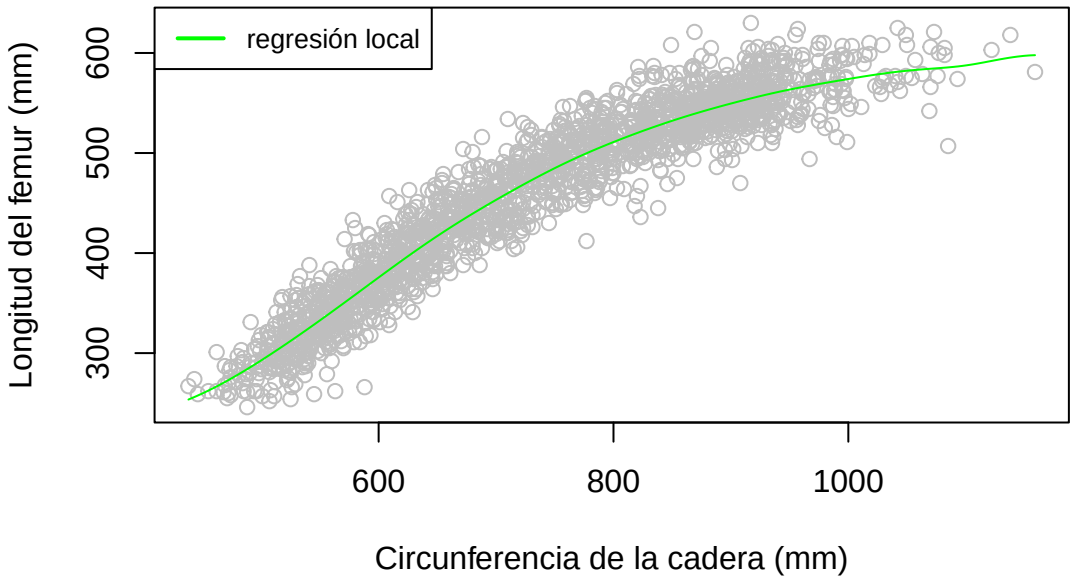
```
## MSE Regresión lineal = 8274.388
```

```
## MSE NW = 929.528
```

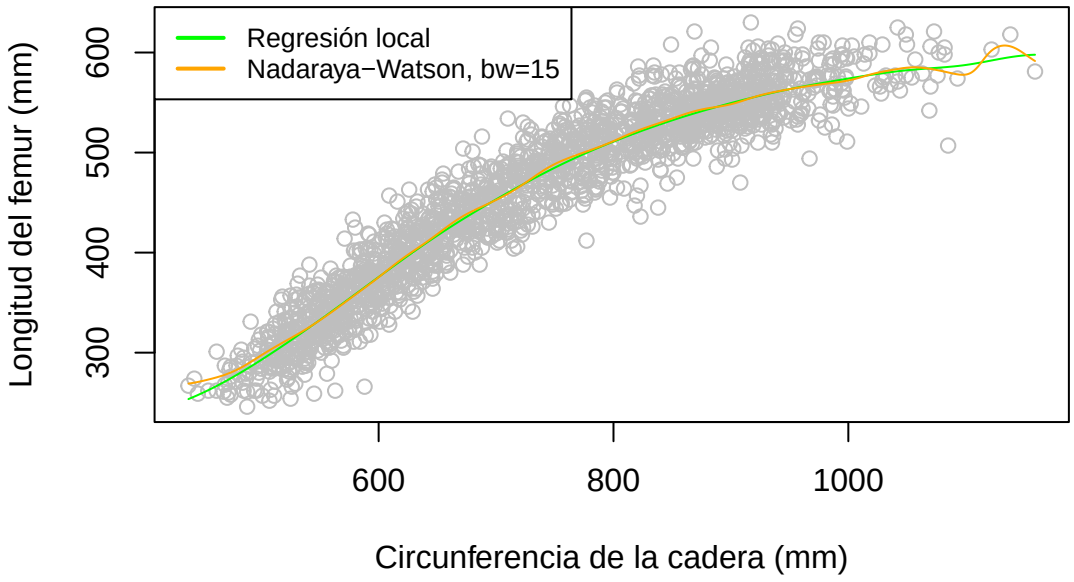
Claramente, nos quedamos con la estimación de NW (tiene menor MSE).

d)

i) y ii)


regresión local
Longitud del femur (mm)
300
400
500
600
600
800
1000
Circunferencia de la cadera (mm)

iii)


Regresión local
Nadaraya−Watson, bw=15
Longitud del femur (mm)
300
400
500
600
600
800
1000
Circunferencia de la cadera (mm)

Podemos ver que la regresión local se ajusta mejor a los datos donde la densidad es baja, mientras que el estimador de Nadaraya-Watson fluctúa más. En cambio, donde hay mayor densidad, ambos estimadores ajustan casi de forma idéntica.