

Trabajo Práctico Final

Introducción a la Estadística y la Ciencia de Datos, 1er Cuatrimestre 2024

Alan Erdei, Nicolás Ian Rozenberg

2024-07-09

1. Sección Teórica

Inciso a)

Demostraremos que el predictor derivado del estimador de Nadaraya-Watson de la esperanza condicional

$$\hat{\mathbf{Y}} = (\hat{Y}_1 \quad \hat{Y}_2 \quad \cdots \quad \hat{Y}_n)^T$$

donde

$$\hat{Y}_i = E(Y|X = X_i) = \hat{m}_h(X_i)$$

es una transformación aplicada sobre \mathbf{Y} .

$$\begin{aligned} \hat{Y}_i &= \hat{m}_h(X_i) = \sum_{j=1}^n y_j w_{j,h}(X_i) \\ &= (w_{1,h}(X_i) \quad w_{2,h}(X_i) \quad \cdots \quad w_{n,h}(X_i)) \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \end{aligned}$$

Por lo tanto

$$\underbrace{\begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix}}_{\hat{\mathbf{Y}}} = \underbrace{\begin{pmatrix} w_{1,h}(X_1) & w_{2,h}(X_1) & \cdots & w_{n,h}(X_1) \\ w_{1,h}(X_2) & w_{2,h}(X_2) & \cdots & w_{n,h}(X_2) \\ \vdots & \vdots & \ddots & \vdots \\ w_{1,h}(X_n) & w_{2,h}(X_n) & \cdots & w_{n,h}(X_n) \end{pmatrix}}_S \underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{\mathbf{Y}}$$

Entonces obtuvimos una transformación lineal en \mathbf{Y} , como se quería ver. ■

2. Sección Práctica

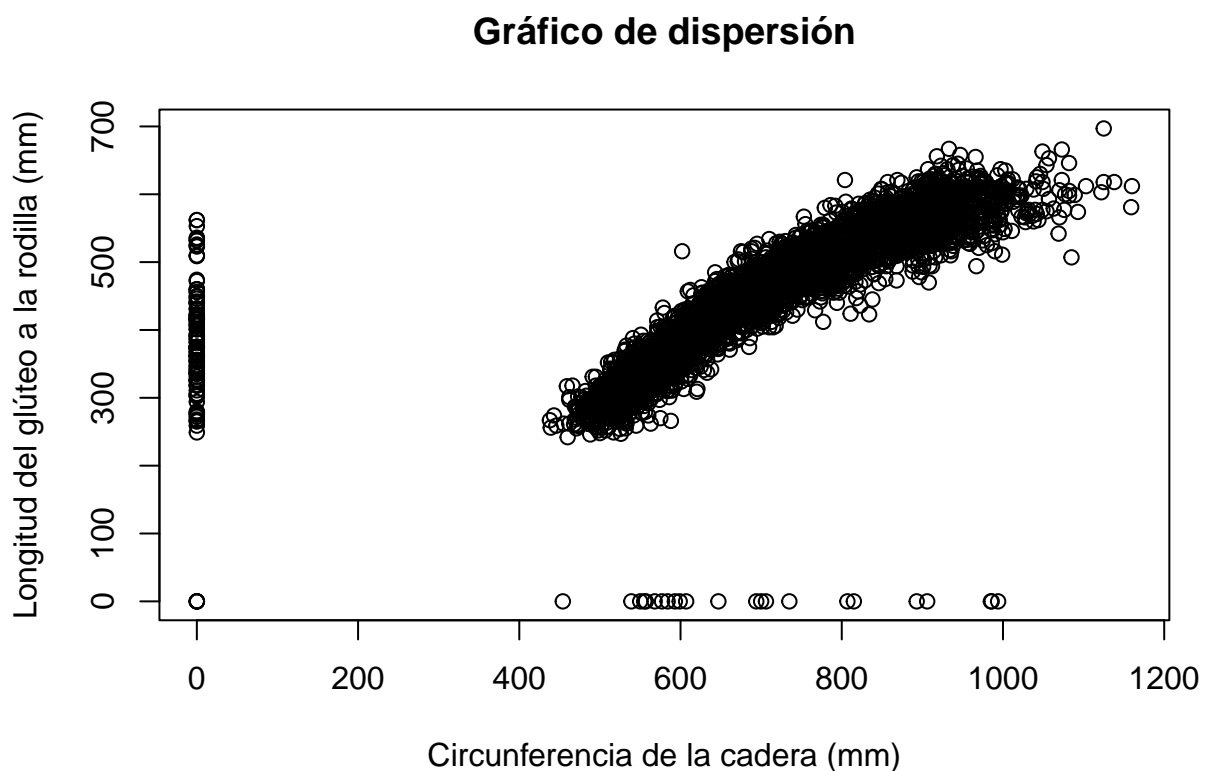
Inciso a)

Cargamos los datos

```
datos <- read.csv("individuals.csv", sep=";")
attach(datos)
```

Primer gráfico de dispersión de Circunferencia de la cadera (HIP.CIRCUMFERENCE) (eje x) vs. Longitud del glúteo a la rodilla (BUTTOCK.KNEE.LENGTH) (eje y)

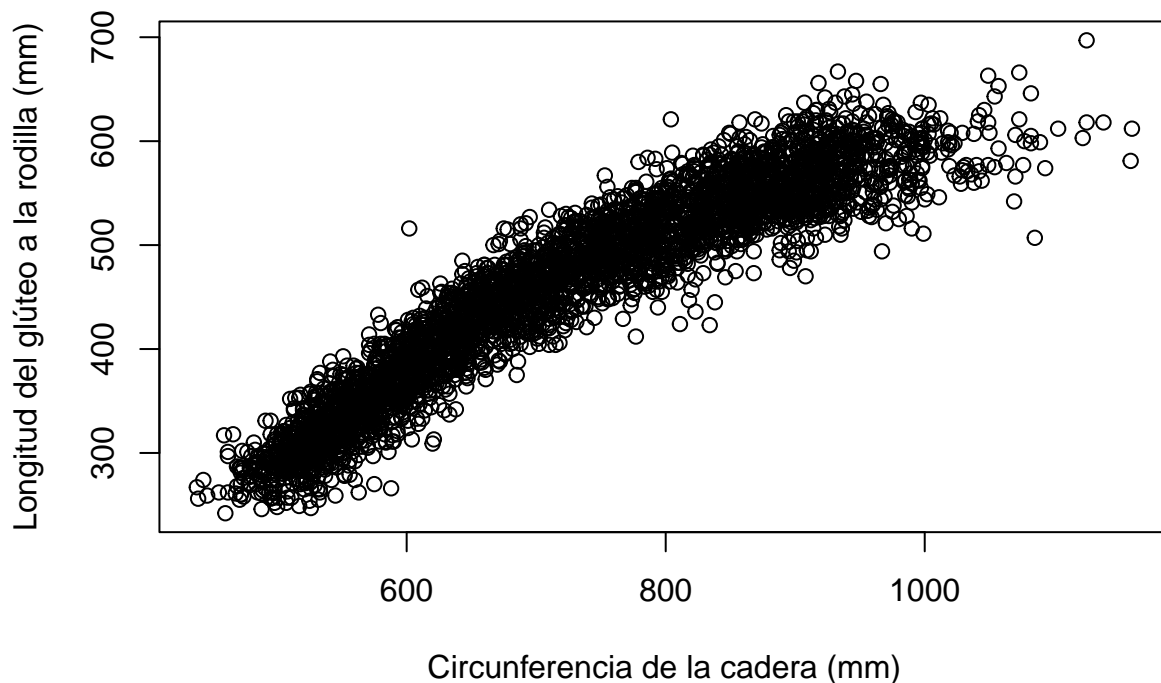
```
plot(
  HIP.CIRCUMFERENCE, BUTTOCK.KNEE.LENGTH,
  main = "Gráfico de dispersión",
  xlab = "Circunferencia de la cadera (mm)",
  ylab = "Longitud del glúteo a la rodilla (mm)"
)
```



Se pueden observar puntos donde o bien la primera variable es 0, o bien la segunda lo es. Podemos interpretar que se desconoce el valor. Graficamos las mismas variables quitando dichos puntos.

```
plot(HIP.CIRCUMFERENCE[HIP.CIRCUMFERENCE != 0 & BUTTOCK.KNEE.LENGTH != 0],
     BUTTOCK.KNEE.LENGTH[HIP.CIRCUMFERENCE != 0 & BUTTOCK.KNEE.LENGTH != 0],
     main = "Gráfico de dispersión",
     xlab = "Circunferencia de la cadera (mm)",
     ylab = "Longitud del glúteo a la rodilla (mm)"
)
```

Gráfico de dispersión



Este gráfico sugiere una correlación positiva entre la circunferencia de la cadera y la longitud del glúteo a la rodilla.

Inciso b

Obtenemos la población femenina

```
poblacion_femenina <- datos[SEX == 2, ]  
  
poblacion_femenina <- poblacion_femenina[  
  poblacion_femenina$HIP.CIRCUMFERENCE != 0 & poblacion_femenina$BUTTOCK.KNEE.LENGTH != 0,  
  ]
```

Separamos a la población femenina en grupos etario de acuerdo al cuartil, y calculamos intervalo de confianza bootstrap de nivel aproximado 0.95. Para cada grupo, generamos un número (`n_bootstrap`, 1000) de muestras bootstrap. Cada muestra bootstrap se obtiene al seleccionar observaciones aleatoriamente con reemplazo del conjunto de datos original del grupo. Esto significa que algunas observaciones pueden ser seleccionadas más de una vez mientras que otras pueden no ser seleccionadas en absoluto. Para cada una de estas muestras bootstrap, calculamos la mediana de `HIP.CIRCUMFERENCE`, y el intervalo de confianza normal. Esto se hizo ordenando las medianas bootstrap y seleccionando los percentiles correspondientes al 2.5% y al 97.5% para un intervalo de confianza del 95%. Estos percentiles representan los límites inferior y superior del intervalo de confianza.

```
quartiles <- quantile(poblacion_femenina$AGE.IN.MONTHS, probs = c(0.25, 0.5, 0.75))  
poblacion_femenina$group <- cut(
```

```

poblacion_femenina$AGE.IN.MONTHS,
breaks = c(-Inf, quartiles, Inf),
labels = c("Q1", "Q2", "Q3", "Q4")
)

ic_bootstrap <- function(data, n_bootstrap = 1000, conf_level = 0.95) {
  medians <- c()

  for (i in 1:n_bootstrap){
    medians <- c(medians, median(sample(data, replace = TRUE)))
  }

  lower_bound <- quantile(medians, (1 - conf_level) / 2)
  upper_bound <- quantile(medians, 1 - (1 - conf_level) / 2)
  list(median = median(data), lower = lower_bound, upper = upper_bound)
}

results <- lapply(
  split(poblacion_femenina$HIP.CIRCUMFERENCE, poblacion_femenina$group), ic_bootstrap
)

for (i in 1:4) {
  cat("Grupo", names(results)[i], "\n")
  cat("Mediana:", results[[i]]$median, "\n")
  cat("Intervalo de confianza:", results[[i]]$lower, "-", results[[i]]$upper, "\n\n")
}

```

```

## Grupo Q1
## Mediana: 556
## Intervalo de confianza: 550.5 - 562
##
## Grupo Q2
## Mediana: 676
## Intervalo de confianza: 669 - 682
##
## Grupo Q3
## Mediana: 799
## Intervalo de confianza: 788 - 808
##
## Grupo Q4
## Mediana: 905
## Intervalo de confianza: 897 - 909.025

```

Graficamos los resultados

```

groups <- names(results)

medians <- sapply(results, function(x) x$median)
lower_bounds <- sapply(results, function(x) x$lower)
upper_bounds <- sapply(results, function(x) x$upper)

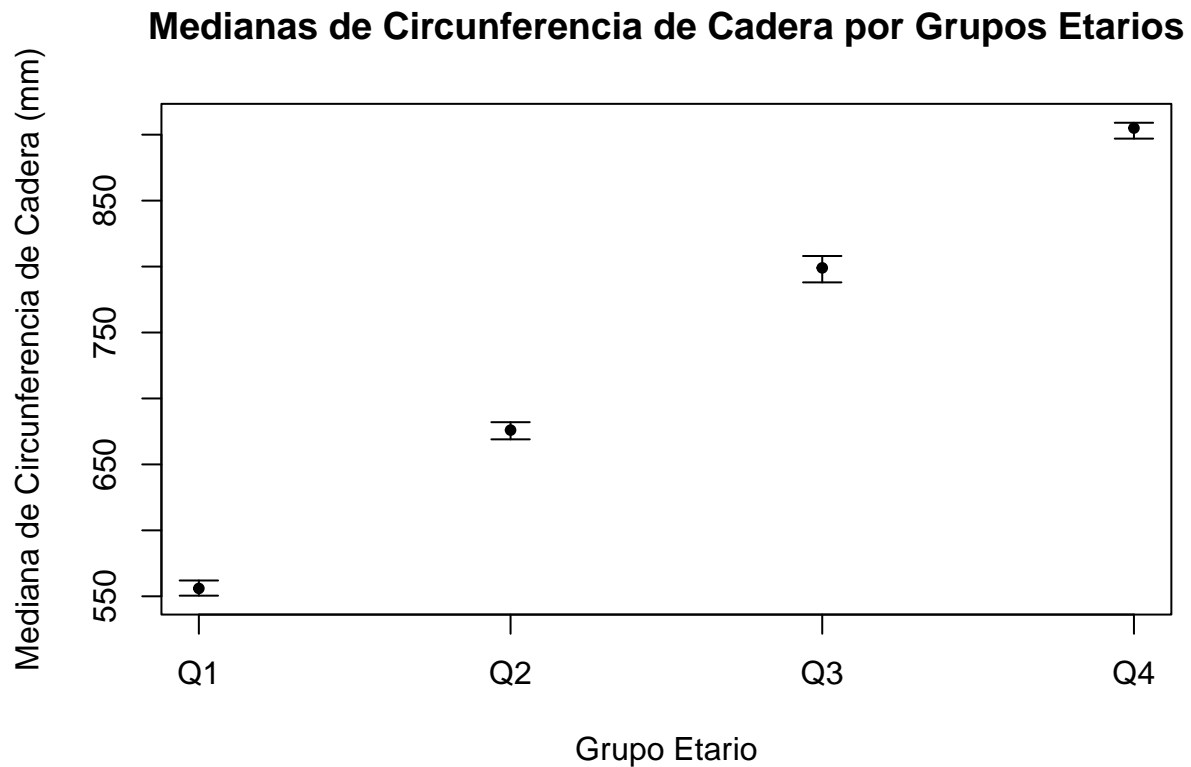
plot(1:4, results$medians, ylim = range(c(lower_bounds, upper_bounds)), xaxt = "n",
     xlab = "Grupo Etario", ylab = "Mediana de Circunferencia de Cadera (mm)",

```

```

main = "Medianas de Circunferencia de Cadera por Grupos Etarios")
axis(1, at = 1:4, labels = groups)
arrows(1:4, lower_bounds, 1:4, upper_bounds, angle = 90, code = 3, length = 0.1)
points(1:4, medians, pch=20)

```



Inciso c)

Agregamos las líneas de regresión al diagrama de dispersión original.

```

X <- HIP.CIRCUMFERENCE[
  HIP.CIRCUMFERENCE != 0 & BUTTOCK.KNEE.LENGTH != 0
]
Y <- BUTTOCK.KNEE.LENGTH[
  HIP.CIRCUMFERENCE != 0 & BUTTOCK.KNEE.LENGTH != 0
]

plot(X, Y,
  main = "Gráfico de dispersión integrando regresión no paramétrica",
  xlab = "Circunferencia de la cadera (mm)",
  ylab = "Longitud del glúteo a la rodilla (mm)"
)
smooth_100 <- ksmooth(X, Y, kernel = "normal", bandwidth = 100)
smooth_50 <- ksmooth(X, Y, kernel = "normal", bandwidth = 50)

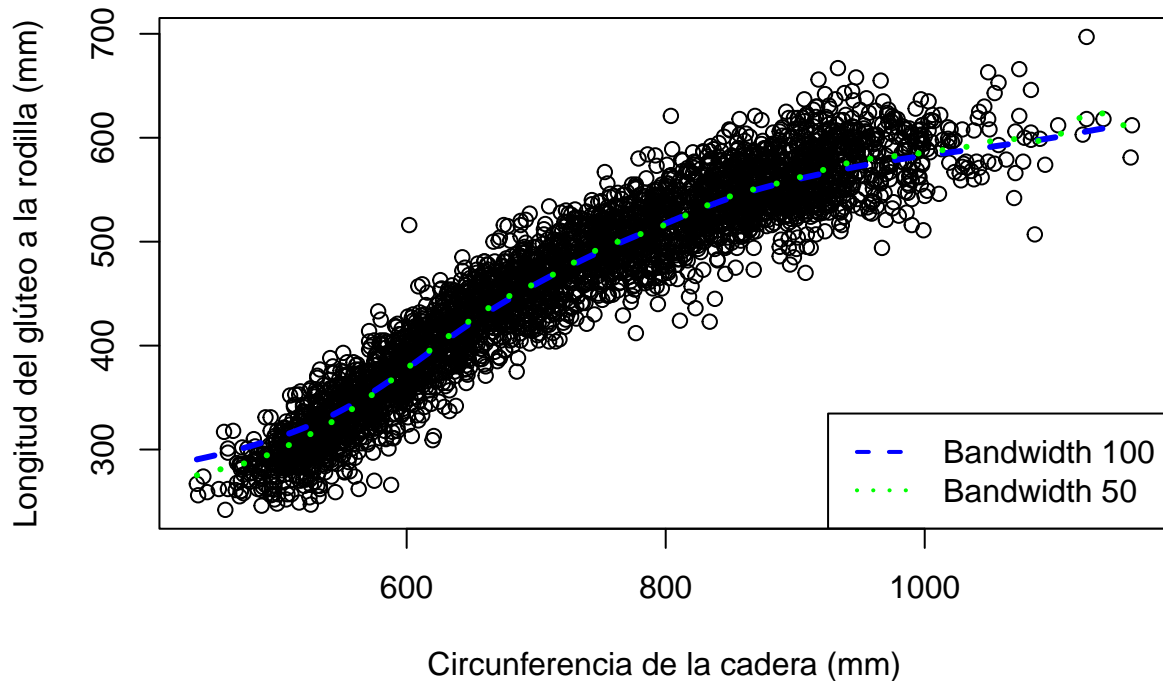
```

```

lines(smooth_100, col = "blue", lwd = 3, lty = 2)
lines(smooth_50, col = "green", lwd = 3, lty = 3)
legend("bottomright", legend = c("Bandwidth 100", "Bandwidth 50"),
      col = c("blue", "green"), lty = c(2, 3), lwd = 2)

```

Gráfico de dispersión integrando regresión no paramétrica



Ahora aplicamos validación cruzada Leave One Out para ksmooth. Por ahora no aplicamos el criterio demostrado en el apartado teórico.

```

loo_mse <- function(bandwidth) {
  n <- length(X)
  mse <- 0

  for (i in 1:n) {
    X_train <- X[-i]
    Y_train <- Y[-i]
    smooth <- ksmooth(X_train, Y_train, kernel = "normal", bandwidth = bandwidth, x.points = X[i])
    mse <- mse + (Y[i] - smooth$y)^2
  }

  return(mse / n)
}

# Búsqueda de la ventana óptima en la grilla de 20 a 50 con paso 1
bandwidths <- 20:50
mse_values <- sapply(bandwidths, loo_mse)

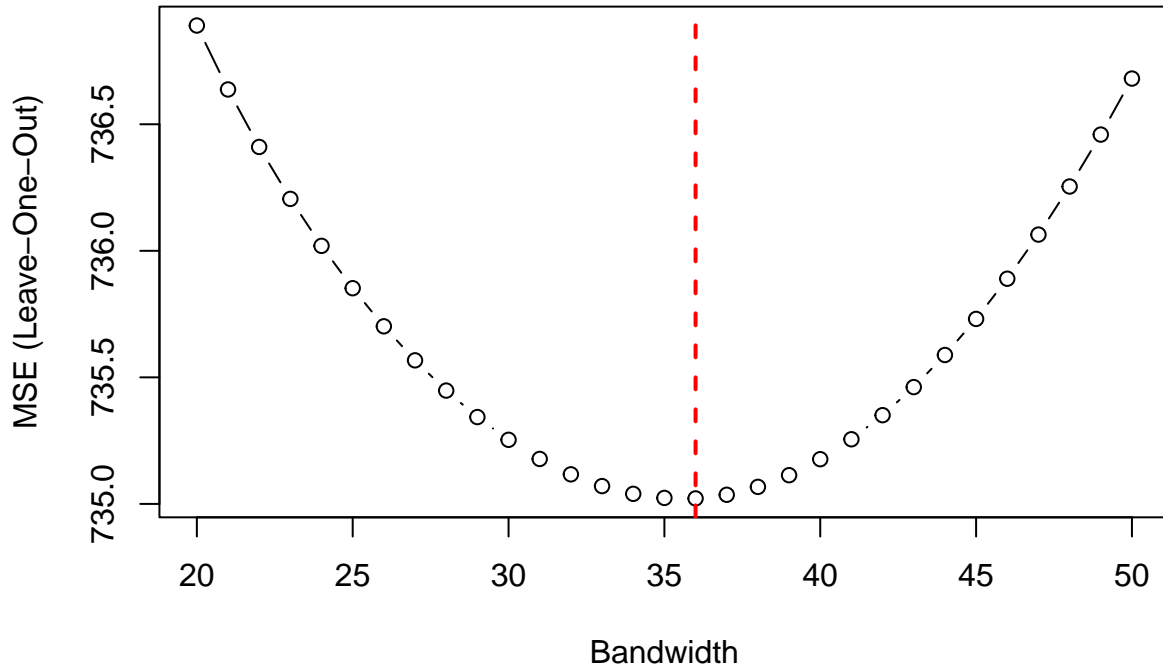
```

```

optimal_bandwidth <- bandwidths[which.min(mse_values)]
plot(bandwidths, mse_values, type = "b", main = "Búsqueda de Ventana Óptima para ksmooth",
     xlab = "Bandwidth", ylab = "MSE (Leave-One-Out)")
abline(v = optimal_bandwidth, col = "red", lwd = 2, lty = 2)

```

Búsqueda de Ventana Óptima para ksmooth



```

cat("La ventana óptima obtenida es", optimal_bandwidth)

```

La ventana óptima obtenida es 36

Ahora, implementamos manualmente el estimador de Nadaraya-Watson con kernel gaussiano

```

# Función para el estimador de Nadaraya-Watson
nwsmooth <- function(x, X, Y, h) {
  K <- function(u) dnorm(u)
  n <- length(X)
  m <- numeric(length(x))

  # Si x == X, weights_matrix corresponde a la matriz asociada con
  # la transformación lineal de Nadaraya-Watson
  weights_matrix <- matrix(nrow = length(x), ncol = n)

  for (i in 1:length(x)) {
    weights_matrix[i, ] <- K((X - x[i]) / h)
    weights_matrix[i, ] <- weights_matrix[i, ] / sum(weights_matrix[i, ])
  }
}

```

```

    m[i] <- sum(weights_matrix[i, ] * Y)
  }

  return(list(m = m, weights = weights_matrix))
}

```

Y ahora realizamos validación cruzada. Ahora sí, utilizando el criterio del apartado teórico.

```

loo_mse_nw <- function(bandwidth) {
  n <- length(X)
  nw_response <- nwsMOOTH(X, X, Y, bandwidth)
  Y_hat <- nw_response$m
  weights <- nw_response$weights
  mse <- 0

  for (i in 1:n){
    mse <- mse + (Y[i]-Y_hat[i])^2 / (1 - weights[i, i])^2
  }

  mse <- mse / n

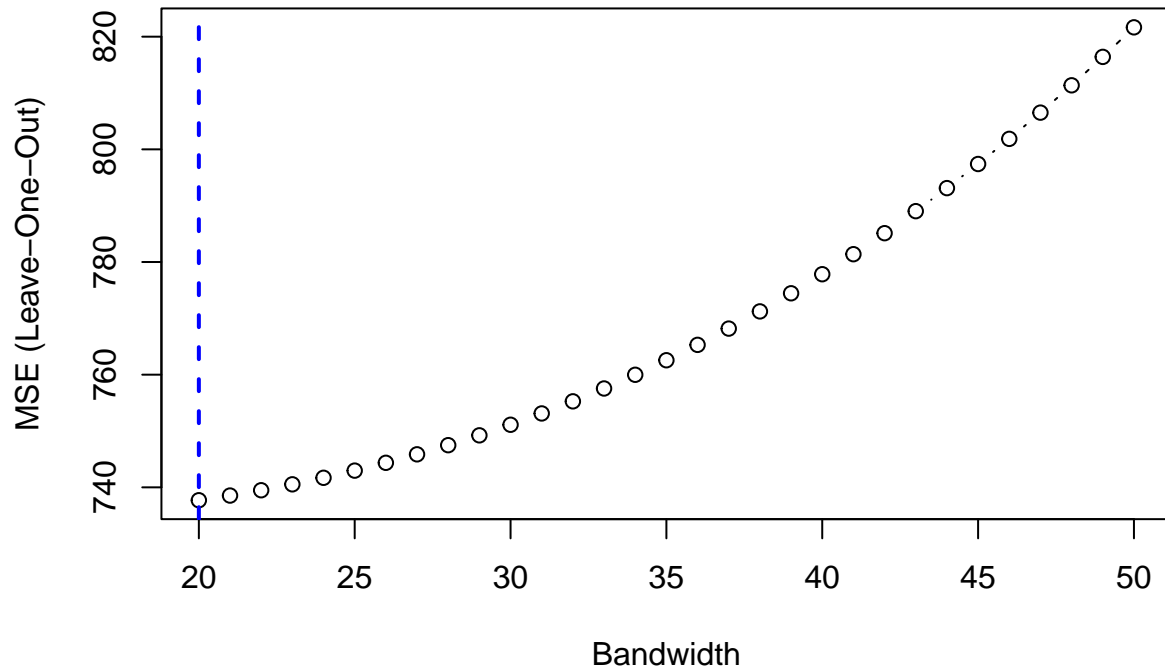
  return(mse)
}

mse_values_nw <- sapply(bandwidths, loo_mse_nw)

optimal_bandwidth_nw <- bandwidths[which.min(mse_values_nw)]
plot(bandwidths, mse_values_nw, type = "b", main = "Búsqueda de Ventana Óptima para Nadaraya-Watson",
     xlab = "Bandwidth", ylab = "MSE (Leave-One-Out)")
abline(v = optimal_bandwidth_nw, col = "blue", lwd = 2, lty = 2)

```


Búsqueda de Ventana Óptima para Nadaraya-Watson

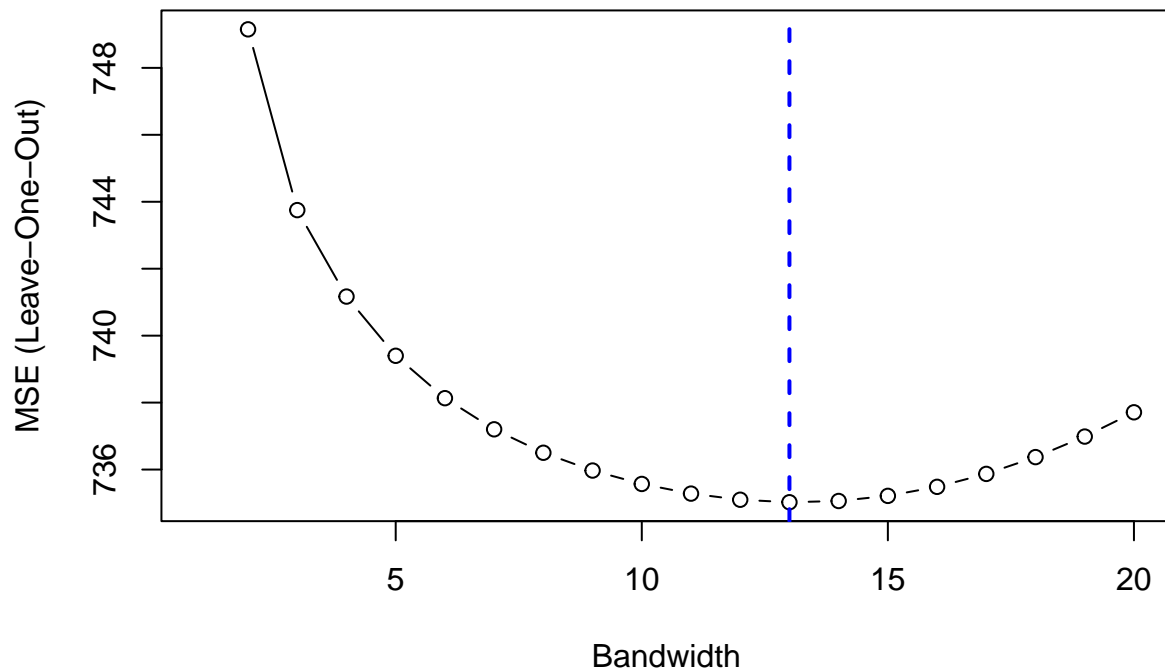


La función objetivo se ve distinta a cuando aplicamos `ksmooth`. Ahora el bandwidth óptimo es un extremo de la grilla. Crearemos una nueva grilla corrida hacia la izquierda, con ventanas que poseen valores que van de 1 a 20

```
bandwidths_nw <- 1:20
mse_values_nw <- sapply(bandwidths_nw, loo_mse_nw)

optimal_bandwidth_nw <- bandwidths_nw[which.min(mse_values_nw)]
plot(bandwidths_nw, mse_values_nw, type = "b", main = "Búsqueda de Ventana Óptima para Nadaraya-Watson",
     xlab = "Bandwidth", ylab = "MSE (Leave-One-Out)")
abline(v = optimal_bandwidth_nw, col = "blue", lwd = 2, lty = 2)
```

Búsqueda de Ventana Óptima para Nadaraya-Watson



Ahora el óptimo en nuestra grilla no se encuentra en un extremo, por lo que podemos quedarnos con dicho valor.

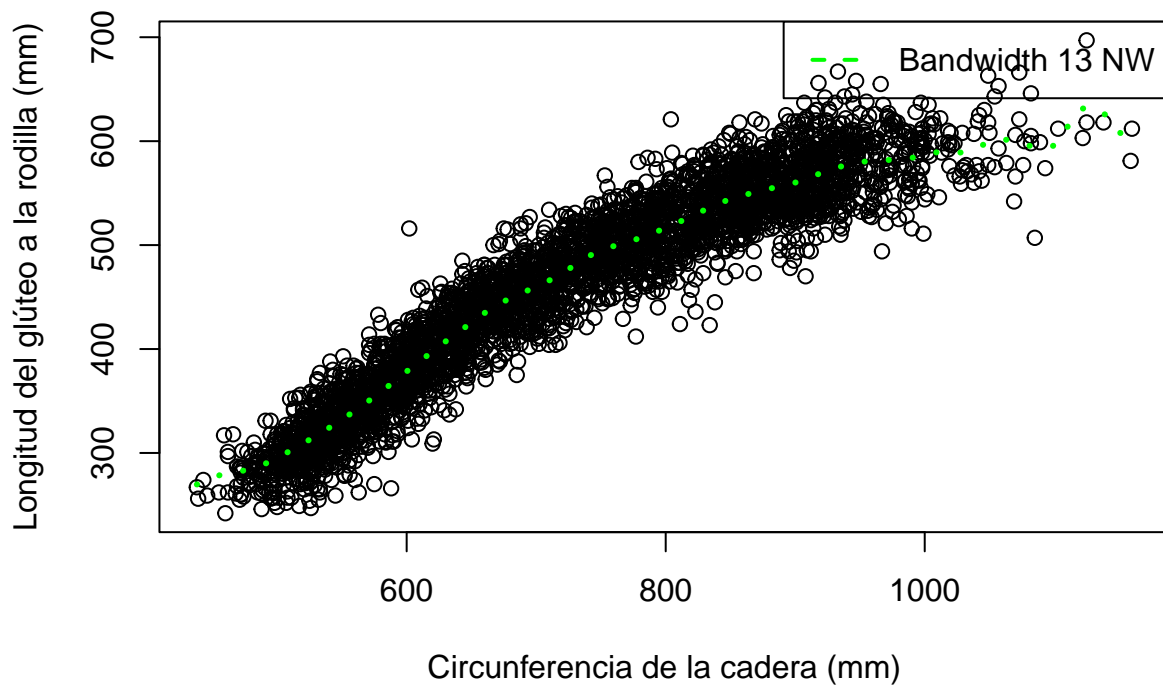
```
cat("La ventana óptima obtenida es", optimal_bandwidth_nw)
```

```
## La ventana óptima obtenida es 13
```

```
plot(X, Y,
     main = "Gráfico de dispersión integrando regresión no paramétrica ksmooth",
     xlab = "Circunferencia de la cadera (mm)",
     ylab = "Longitud del glúteo a la rodilla (mm)"
)

x <- seq(min(X), max(X), length.out = 2*length(X))
smooth_nw_opt <- nwsMOOTH(x, X, Y, optimal_bandwidth_nw)
lines(x, smooth_nw_opt$m, col = "green", lwd = 3, lty = 3)
legend("topright", legend = sprintf("Bandwidth %s NW", optimal_bandwidth_nw),
      col = "green", lty = 2, lwd = 2)
```

Gráfico de dispersión integrando regresión no paramétrica ksmooth



Ahora nos falta realizar un gráfico viendo cómo queda la regresión obtenida por cuadrados mínimos, y por las mejores ventanas tanto de `ksmooth` como de `nwsmooth` (iv)

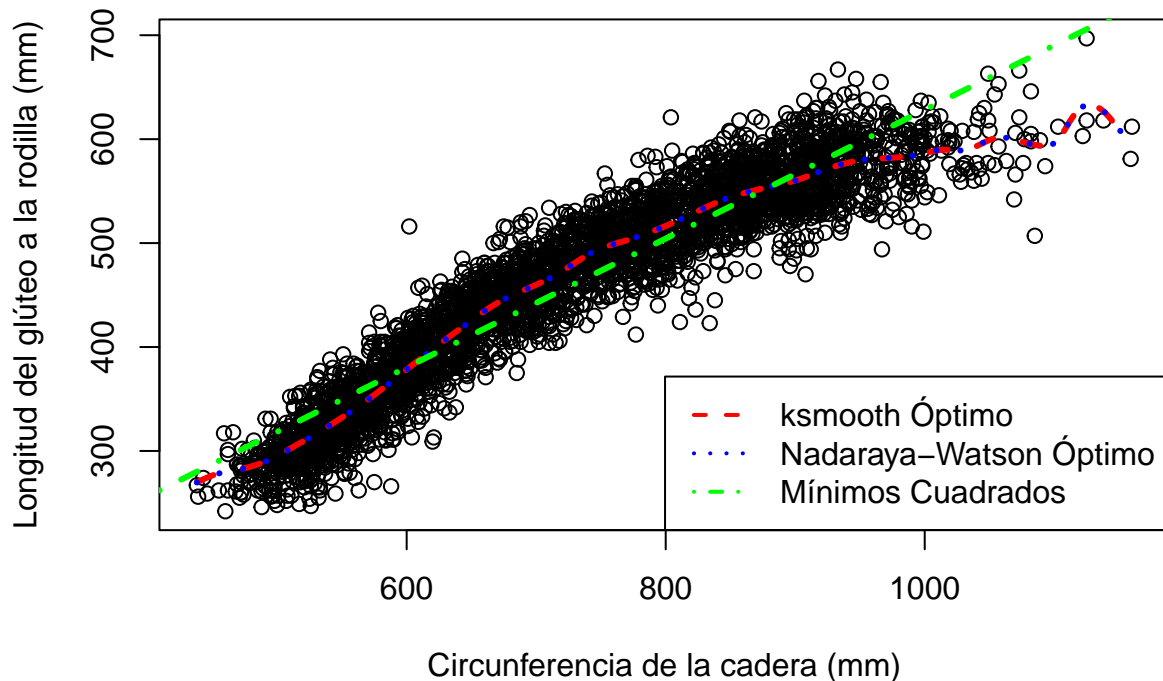
```
smooth_opt <- ksmooth(X, Y, kernel = "normal", bandwidth = optimal_bandwidth)

fit_lm <- lm(Y ~ X)

plot(X, Y,
     main = "Gráfico de dispersión integrando regresión no paramétrica nwsmooth",
     xlab = "Circunferencia de la cadera (mm)",
     ylab = "Longitud del glúteo a la rodilla (mm)"
)

lines(smooth_opt, col = "red", lwd = 3, lty = 2)
lines(x, smooth_nw_opt$m, col = "blue", lwd = 3, lty = 3)
abline(fit_lm, col = "green", lwd = 3, lty = 4)
legend("bottomright", legend = c("ksmooth Óptimo", "Nadaraya-Watson Óptimo", "Mínimos Cuadrados"),
     col = c("red", "blue", "green"), lty = c(2, 3, 4), lwd = 2)
```

Gráfico de dispersión integrando regresión no paramétrica nwsmo



Observamos que `ksmooth` y `nwsmooth` se comportan de forma prácticamente idéntica en sus ventanas óptimas. Sólo observando el gráfico, consideramos que no se puede determinar cuál de los dos enfoques preferiríamos. Considerando que existe cierto cambio en la correlación entre las dos variables en valores altos de circunferencia de la cadera, el modelo no paramétrico podría servir para predecir mejor dichas situaciones. Sin embargo, a medida que crece el valor de las variables, los datos se vuelven más escasos. Entonces la desventaja del enfoque no paramétrico es que necesita de una cantidad mayor de datos para ser estable que en el enfoque paramétrico dado por el modelo lineal. Calculemos los R^2 (coeficiente de determinación)

```
cat("R^2 modelo lineal", summary(fit_lm)$r.squared)
```

```
## R^2 modelo lineal 0.8856531
```

```
calcular_r2 <- function(Y_real, Y_hat_ksmooth){  
  ss_total <- sum((Y_real - mean(Y))^2)  
  ss_res_ksmooth <- sum((Y_real - Y_hat_ksmooth)^2)  
  rsq <- 1 - (ss_res_ksmooth / ss_total)  
  
  return(rsq)  
}
```

```
smooth_nw_opt <- nwsmooth(X, X, Y, optimal_bandwidth_nw)  
cat("R^2 NW", calcular_r2(Y, smooth_nw_opt$m))
```

```
## R^2 NW 0.9211362
```

```
cat("R^2 ksmooth", calcular_r2(sort(Y), smooth_opt$y))
```

```
## R^2 ksmooth 0.8759315
```

El método que mejor score R^2 tuvo es el enfoque no paramétrico con el estimador de Nadaraya Watson implementado manualmente. Por lo tanto, es el que mejor explica la variación en los datos. Sin embargo, hay que tener en consideración las observaciones previas.

Inciso d)

Implementaremos dicho predictor basándonos en las cuentas que se encuentran en Loader (2004, p. 4-5).

```
linearsmooth <- function(x, X, Y, h){
  K <- function(u) dnorm(u)
  n <- length(X)
  m <- numeric(length(x))

  weights_matrix <- matrix(nrow = n, ncol = n)

  for (i in 1:n) {
    weights_matrix[i, ] <- K((X - X[i]) / h)
    weights_matrix[i, ] <- weights_matrix[i, ] / sum(weights_matrix[i, ])
  }

  for (i in 1:length(x)) {
    X_design <- cbind(rep(1, n), X - x[i])
    M1 <- solve(t(X_design) %*% weights_matrix %*% X_design)
    M2 <- t(X_design) %*% weights_matrix
    coefs <- M1 %*% M2 %*% Y

    m[i] <- coefs[1]
  }

  return(m)
}

Y_pred <- linearsmooth(seq(min(X), max(X), length.out=500), X, Y, 13)

plot(X, Y,
     main = "Gráfico de dispersión integrando regresión no paramétrica nwsmooth",
     xlab = "Circunferencia de la cadera (mm)",
     ylab = "Longitud del glúteo a la rodilla (mm)"
)

lines(seq(min(X), max(X), length.out=500), Y_pred, col = "blue", lwd = 3, lty = 3)
```

Gráfico de dispersión integrando regresión no paramétrica nws moo

