Trabajo Práctico Final

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

Alan Erdei, Nicolás Ian Rozenberg

2024-07-06

1. Sección Teórica

Inciso a)

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

$$\hat{\boldsymbol{Y}} = \begin{pmatrix} \hat{Y_1} & \hat{Y_2} & \cdots & \hat{Y_n} \end{pmatrix}^T$$

donde

$$\hat{Y_i} = E(Y|X = X_i) = \hat{m_h}(X_i)$$

es una transformación aplicada sobre $\mathbf{Y}$.

$$\hat{Y_i} = \hat{m_h}(X_i) = \sum_{j=1}^{n} y_j w_{j,h}(X_i)$$

$$= \begin{pmatrix} w_{1,h}(X_i) & w_{2,h}(X_i) & \cdots & w_{n,h}(X_i) \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

Por lo tanto

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

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

2. Sección Práctica

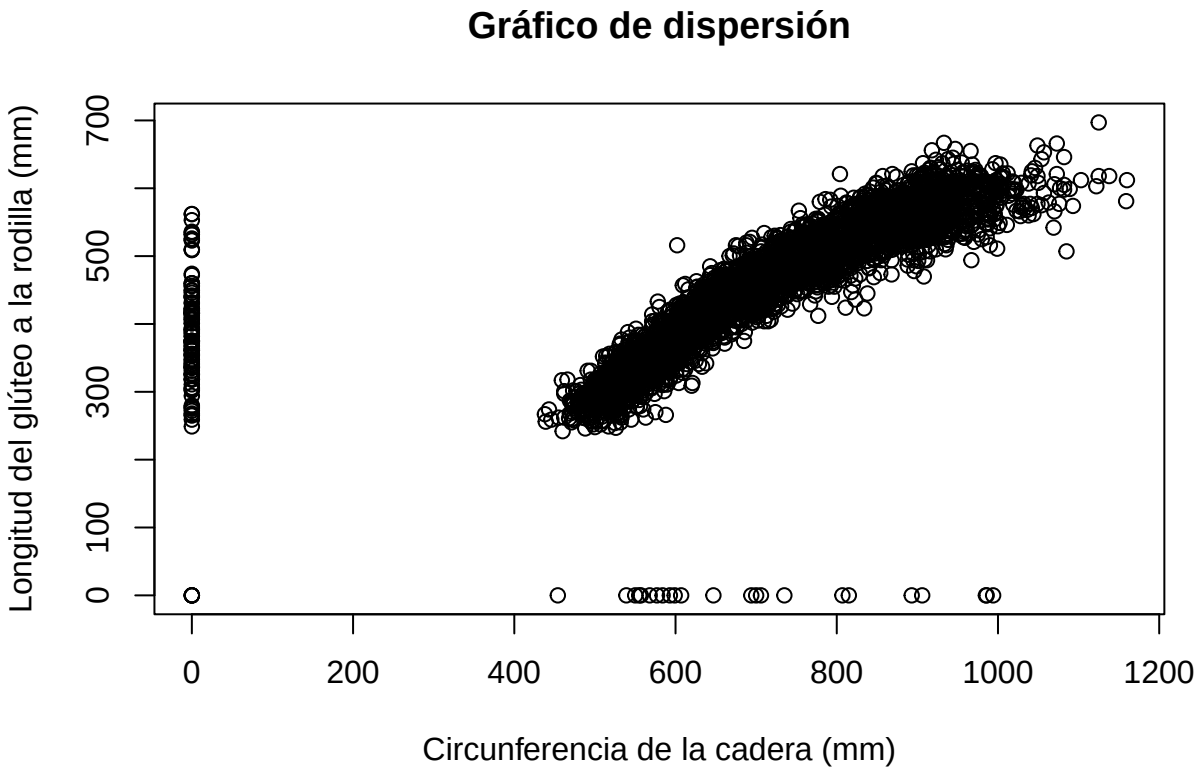
Inciso a)

Cargamos los datos

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

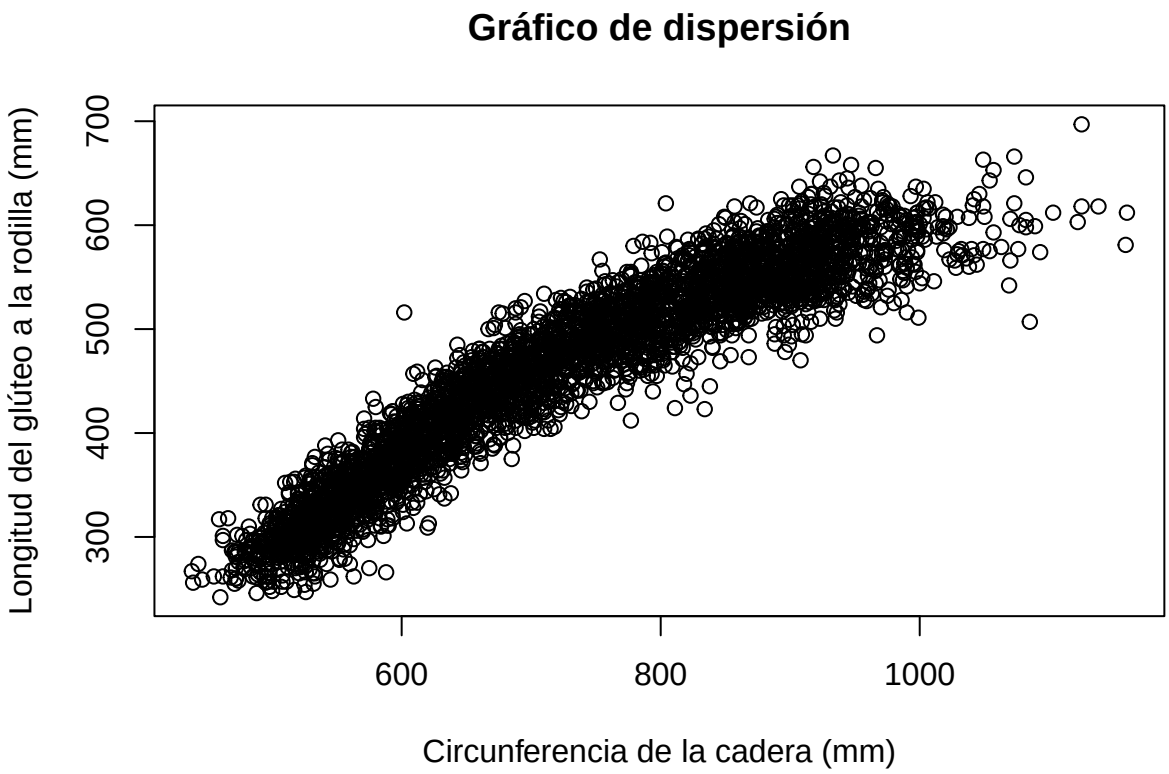
Primer gráfico de dispersión de Circunferencia de la cadera (HIP.CIRCUMFERENCE) (eje x) vs. Longitud del glúteo a la rodilla (BUTTOCK.KNEE.LENGTH) (eje y)

```r
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.

```r
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)"
)
```



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 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), labe
```

```r
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")
}
```

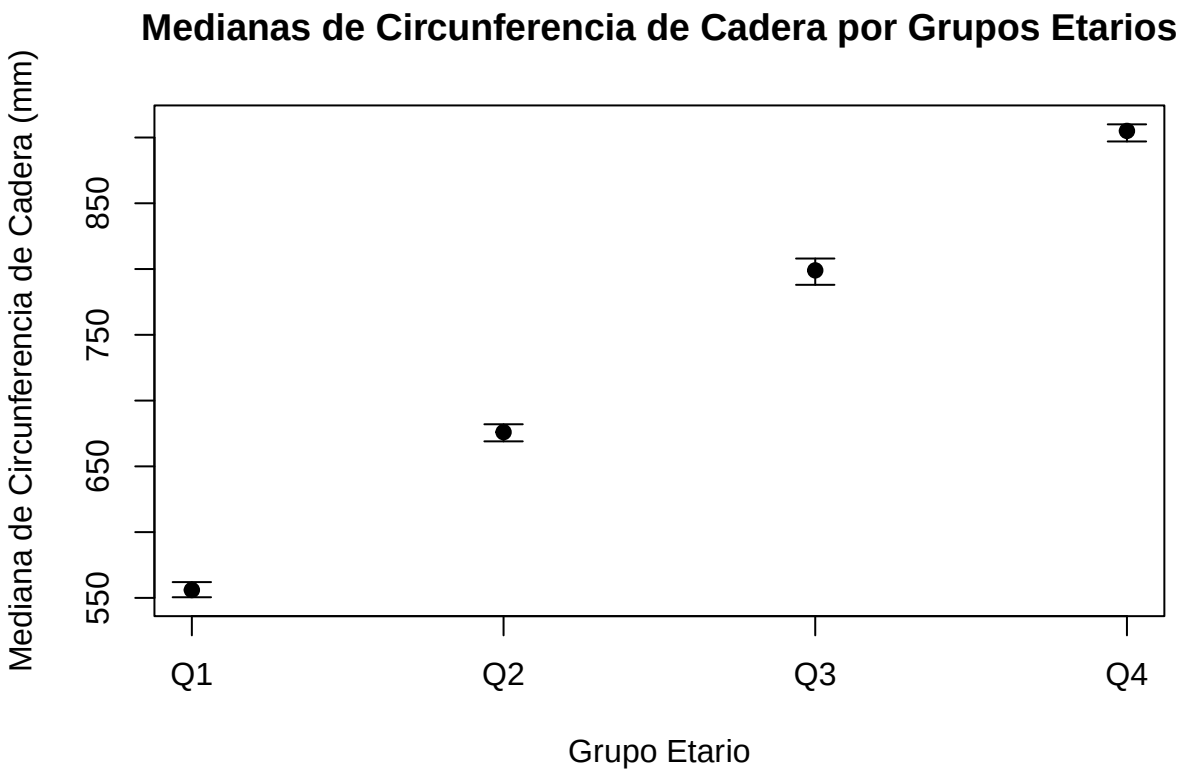
```
## Grupo Q1
## Mediana: 556
## Intervalo de confianza: 550.5 - 562
##
## Grupo Q2
## Mediana: 676
## Intervalo de confianza: 669 - 682
##
## Grupo Q3
## Mediana: 799
## Intervalo de confianza: 788 - 808
##
## Grupo Q4
## Mediana: 905
## Intervalo de confianza: 897 - 910
```

Graficamos los resultados

```r
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 = 19)
```


Medianas de Circunferencia de Cadera por Grupos Etarios
Mediana de Circunferencia de Cadera (mm)
550
650
750
850
Q1
Q2
Q3
Q4
Grupo Etario