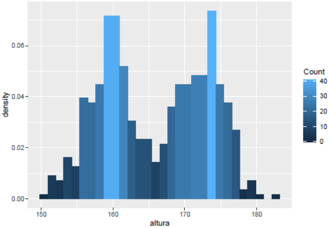
Regresión No Paramétrica

Ana M. Bianco & Paula M. Spano

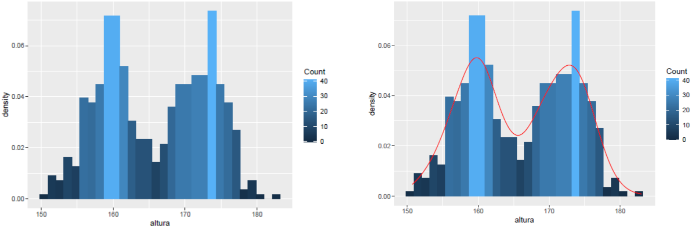
Introducción al Aprendizaje Estadístico

Primera Parte
Volvamos a los datos de altura del hije

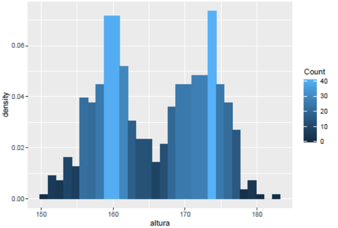
Y = altura hije

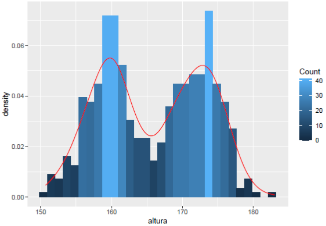


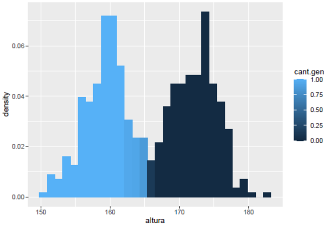
Y = altura hije



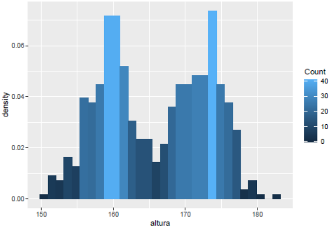
Y = altura hije

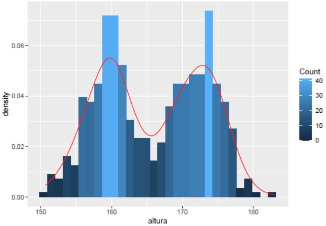


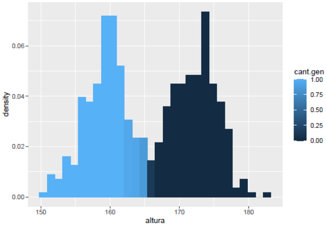


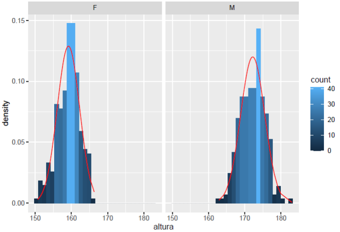


Y = altura hije

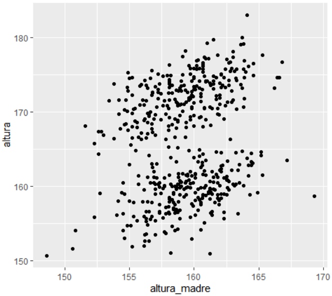




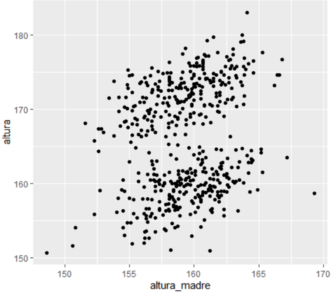


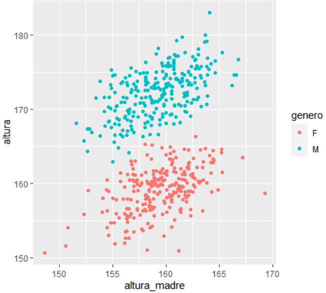


Y = altura hije vs. X =altura madre



Y = altura hije vs. X =altura madre





Mínimos cuadrados

Vayamos al shiny.

Regresión: Galton

ANTHROPOLOGICAL MISCELLANEA.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.

[WITH PLATES IX AND X.]

¿Cuánto medirá de grande?

¿Cuánto medirá de grande?

¿Qué información adicional tenemos?

Escenario 1: ninguna

Escenario 2: sexo

sabemos algo de la madre....

Escenario 3 : será varón y la madre es bajita.

Escenario 4 : será varón y la madre mide 156 cm.

Paso a Paso

La pregunta:¿Cuánto medirá de grande?

Escenario 1: SIN información adicional

Tenemos datos. ¿Qué hacemos? 🤔

Paso a Paso

La pregunta:¿Cuánto medirá de grande?

Escenario 1: SIN información adicional

Tenemos datos. ¿Qué hacemos? 🤔

Promediamos!!

Paso a Paso

La pregunta:¿Cuánto medirá de grande?

Escenario 2: Será varón

Tenemos datos. ¿Qué hacemos? 🤔

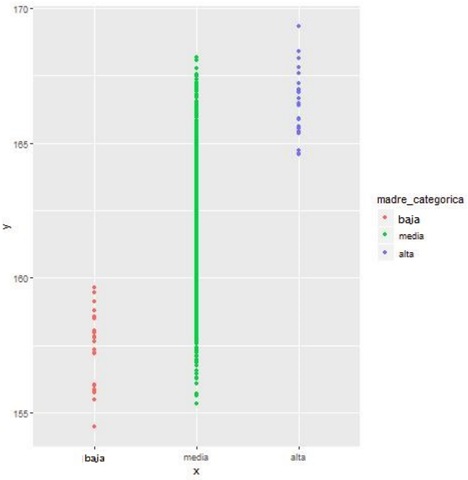
Paso a Paso

La pregunta:¿Cuánto medirá de grande?

Escenario 3: Será varón y la mamá es bajita.

Tenemos datos. ¿Qué hacemos? 🤔

Y vs. X: categórica



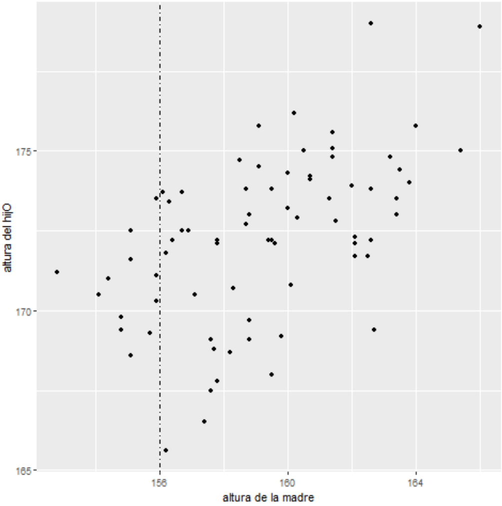
Paso a Paso

La pregunta:¿Cuánto medirá de grande?

Escenario 4: Será varón y la mamá mide 156cm.

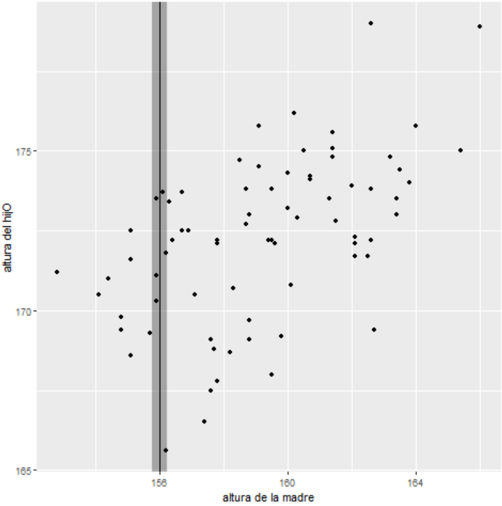
Tenemos datos. ¿Qué hacemos? 🤔

La madre mide 156



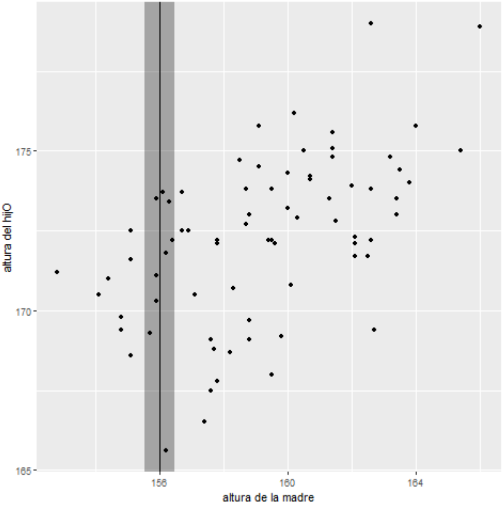
¿Podemos predecir la altura del hijo de esta mamá?

Abrimos una ventana - Opción 1



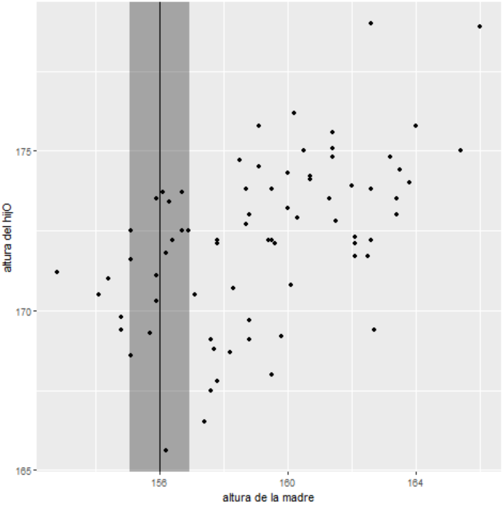
Abrimos una ventana y promediamos las alturas de los hijos correspondientes a los pares que caen adentro.

Ensanchamos la ventana - Opción 2



Abrimos más la ventana y promediamos las alturas de los hijos de los pares que caen adentro.

Ensanchamos aún más la ventana - Opción 3



Ensanchamos más aún la ventana y promediamos las alturas de los hijos de los pares que caen adentro.

Vayamos a los ítem 9 a 11 de las Tareas de Clase

¿Qué proponemos? Predecir con promedios locales

1. Determinamos la altura x de la madre donde quiere predecir (156)

2. Elegimos un valor h de ventana para armar la vecindad (1)

3. Promediamos las alturas de los hijos correspondientes a los pares que caen adentro de la vecindad con ventana de tamaño h ($\pm h$) centrada en x.

Vayamos al shiny a experimentar un poco

El jardín de senderos que se bifurcan...

- Usar distintos pesos dentro de la vecindad: núcleos

- Calcular medianas en lugar de promedios: medianas locales

- Elegir a las k-madres más cercanas y promediar la altura de sus hijos.

- Elegir a las k-madres mas cercanas y calcular la mediana de la altura de sus hijos.

- Nadaraya-Watson (estimador de núcleos).
- Vecinos más cercanos (knn).

Se va la segunda...

Estimación no paramétrica de la regresión

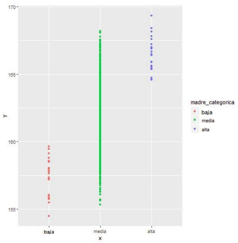
- Modelo $Y = r(X) + \varepsilon$; X, ε independientes, $\mathbb{E}(\varepsilon) = 0$
- Función de regresión: $r(x) = \mathbb{E}(Y \mid X = x)$.

Estimación no paramétrica de la regresión

- Modelo $Y = r(X) + \varepsilon$; X, ε independientes, $\mathbb{E}(\varepsilon) = 0$
- Función de regresión: $r(x) = \mathbb{E}(Y \mid X = x)$.

- Muestra: $\{(X_i, Y_i) : i = 1, \ldots, n\}$
- Estimación de $r(\mathbf{x})$ - Caso X discreta.

$$\widehat{r}_n(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i \, I_{\{X_i = x\}}}{\sum_{i=1}^{n} I_{\{X_i = x\}}}$$



Estimación no paramétrica de la regresión

- Modelo $Y = r(X) + \varepsilon$; X, ε independientes, $\mathbb{E}(\varepsilon) = 0$
- Función de regresión: $r(x) = \mathbb{E}(Y \mid X = x)$.

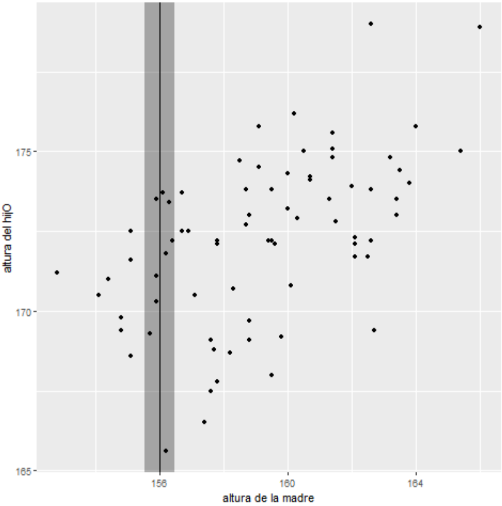
- Muestra: $\{(X_i, Y_i) : i = 1, \ldots, n\}$
- Estimación de $r(\mathbf{x})$ - Caso X discreta.

$$\widehat{r}_n(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i \; I_{\{X_i = x\}}}{\sum_{i=1}^n I_{\{X_i = x\}}}$$

$$\widehat{r}_n(x) = \frac{\sum_{i=1}^n Y_i \; I_{\{|X_i - x| = 0\}}}{\sum_{i=1}^n I_{\{|X_i - x| = 0\}}}$$

- ¿ y si X continua?

X continua



Estimador de Núcleos de Nadaraya - Watson

- Estimación de $r(x) = \mathbb{E}(Y \mid X = x)$ - X continua.

$$\widehat{r}_n(x) = \frac{\sum_{i=1}^{n} Y_i I_{\{|X_i - x| \leq h\}}}{\sum_{i=1}^{n} I_{\{|X_i - x| \leq h\}}}$$

$$\widehat{r}_n(x) = \frac{\sum_{i=1}^{n} Y_i I_{\left\{|\frac{X_i - x}{h}| \leq 1\right\}}}{\sum_{i=1}^{n} I_{\left\{|\frac{X_i - x}{h}| \leq 1\right\}}}$$

Estimador de Núcleos de Nadaraya - Watson

- Estimación de $r(x) = \mathbb{E}(Y \mid X = x)$ - X continua.

$$\begin{aligned}
\widehat{r}_n(x) &= \frac{\sum_{i=1}^{n} Y_i I_{\{|X_i - x| \leq h\}}}{\sum_{i=1}^{n} I_{\{|X_i - x| \leq h\}}} \\
\widehat{r}_n(x) &= \frac{\sum_{i=1}^{n} Y_i I_{\left\{\left|\frac{X_i - x}{h}\right| \leq 1\right\}}}{\sum_{i=1}^{n} I_{\left\{\left|\frac{X_i - x}{h}\right| \leq 1\right\}}} \\
K(u) &= \frac{1}{2} I_{|u| \leq 1}\, , K = f_U, U \sim \mathcal{U}[-1, 1] \\
\widehat{r}_n(x) &= \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}
\end{aligned}$$

Estimador de Núcleos de Nadaraya–Watson

$$\widehat{r}_n(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}$$

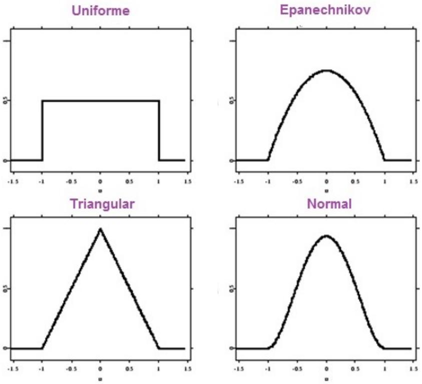
$$\frac{\displaystyle\sum_{i=1}^n Y_i \, K\left(\frac{X_i - x}{h}\right)}{\displaystyle\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)} = \sum_{i=1}^n Y_i \, \underbrace{\frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}}_{\substack{\shortparallel \\ W_i}}$$

$$\begin{aligned} \widehat{r}_n(x) &= \sum_{i=1}^n Y_i \, W_i(x) \\ &= \text{promedio ponderado por la distancia a } x. \end{aligned}$$

Tipos de núcleos

- Núcleo Rectangular: $K(t) = \frac{1}{2}\mathcal{I}_{[-1,1]}(t)$
- Núcleo Triangular: $K(t) = (1 - |t|)\mathcal{I}_{[-1,1]}(t)$
- Núcleo Gausssiano: $K(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}$
- Núcleo Epanechnikov: $K(t) = \frac{3}{4}(1 - t^2)\mathcal{I}_{[-1,1]}(t)$

Núcleos



Estimadores de Nadaraya–Watson

Proponemos estimadores de la forma:

$$\widehat{r}_n(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}$$

$K(t)$ es un kernel y $h > 0$ es el ancho de ventana, donde $K(t)$ satisface:

i) $K(t) \geq 0$

ii) $K(t) = K(-t)$ (función par)

iii) $\int K(t)dt = 1$

iv) $\int tK(t)dt = 0$

v) $\int t^2 K(t)dt < \infty$

.

knn- Vecinos más cercanos - Stone (1977)

Promediamos las respuestas de los k vecinos que están más cerca en el espacio de la covariable.

knn- Vecinos más cercanos - Stone (1977)

Promediamos las respuestas de los k vecinos que están más cerca en el espacio de la covariable.

- Ordenamos X_i según la distancia a x.

$$||X^{(1)} - x|| < ||X^{(2)} - x|| < \ldots < ||X^{(n)} - x||$$

- d_x^k=distancia de x al k-ésimo vecino más cercano: $||X^{(k)} - x||$.
- Entorno con los k- más cercanos.

$$\mathcal{E}_{\mathbf{x}} = \{i \in \{1, \ldots, n\} : ||X_i - x|| \leq d_x^k\}$$

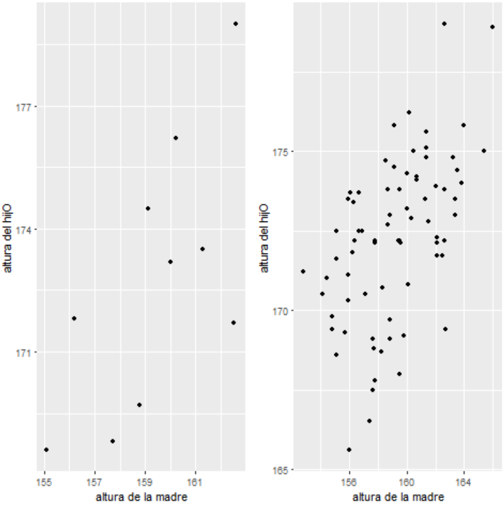
$$\widehat{r}(x) = \widehat{r}_k(x) = \frac{1}{k} \sum_{i \in \mathcal{E}_{\mathbf{x}}} Y_i$$

Volvamos a las Tareas de Clase

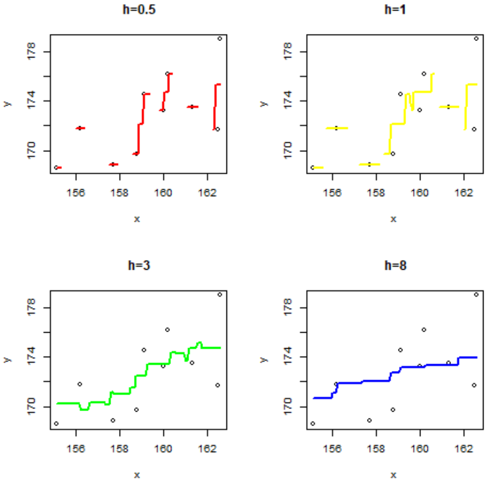
Regresión no paramétrica

Vayamos al shiny.

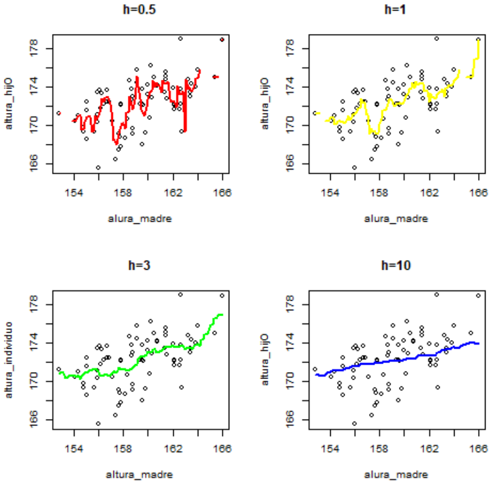
Pocos Datos vs. Muchos Datos



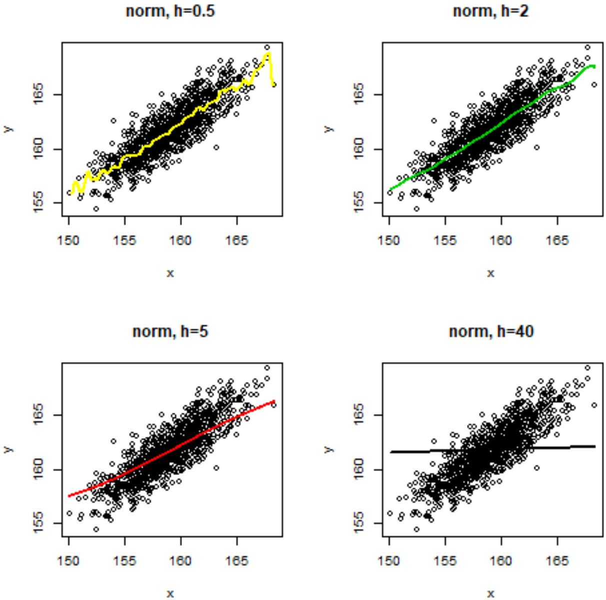
Promedios locales con pocos datos



Promedios locales con *muchos* datos



Efecto Ventana



¿Cómo elegimos los parámetros de suavizado?

Tuning Parameter: h y k

- Nadaraya-Watson. ventana: h. $\widehat{r}_h(x)$
- Vecinos más cercanos vecinos (knn). vecinos: k. $\widehat{r}_k(x)$
- Caso general: $\widehat{r}_t(x)$. tunning parameter: t:

Test Error vs. Training Error

- $\mathcal{M} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$
  $\widehat{r} = \widehat{r}_{t,\mathcal{M}} = \widehat{r}_n$. Predicción: $\widehat{r}_{t,\mathcal{M}}(X)$

Test Error vs. Training Error

- $\mathcal{M} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$
  $\widehat{r} = \widehat{r}_{t,\mathcal{M}} = \widehat{r}_n$. Predicción: $\widehat{r}_{t,\mathcal{M}}(X)$
- $\mathbb{E}\left\{[Y_{\text{new}} - \widehat{r}_{t,\mathcal{M}}(X_{\text{new}})]^2 \mid \mathcal{M}\right\}$

Test Error vs. Training Error

- $\mathcal{M} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$
  $\widehat{r} = \widehat{r}_{t,\mathcal{M}} = \widehat{r}_n$. Predicción: $\widehat{r}_{t,\mathcal{M}}(X)$
- $\mathbb{E}\left\{[Y_{\text{new}} - \widehat{r}_{t,\mathcal{M}}(X_{\text{new}})]^2 \mid \mathcal{M}\right\}$ Test error
- $i-$ésimo Error de Predicción:

$$Y_i - \widehat{r}_{t,\mathcal{M}}(X_i)$$

- Error Cuadrático de Predicción Promediado

$$ECPP(h) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \widehat{r}_{t,\mathcal{M}}(X_i)\}^2$$

Splitting the data $\mathcal{M}$

- $\mathcal{T}$: Muestra de entrenamiento: es usada para ajustar el modelo
- $\mathcal{V}$: Muestra de validación: es usada para seleccionar el modelo

Cuando hay muchos datos

- $\mathcal{T}$: Muestra de entrenamiento: es usada para ajustar el modelo. 80%:
- $\mathcal{V}$: Muestra de validación: es usada para seleccionar el modelo. 20%

$$\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}(Y_i - \widehat{r}_{t,\mathcal{T}}(\mathbf{X}_i))^2$$

Cuando hay muchos datos

- $\mathcal{T}$: Muestra de entrenamiento: es usada para ajustar el modelo. 80%:
- $\mathcal{V}$: Muestra de validación: es usada para seleccionar el modelo. 20%

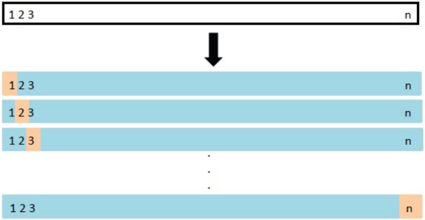
$$\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\left(Y_i - \widehat{r}_{t,\mathcal{T}}(\mathbf{X}_i)\right)^2$$

$$t_{opt} = \textit{argmin}\,\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\left(Y_i - \widehat{r}_{t,\mathcal{T}}(\mathbf{X}_i)\right)^2$$

Menos Datos- Selección del Tunning Parameter

Cross Validation

Cross Validation: Leave one out - Representación esquemática (ISLR)



Cross Validation: Leave one out - Fórmulas

t: tunning parameter

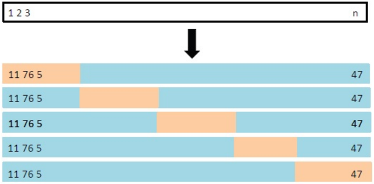
$$\mathsf{CV}(t) = \frac{1}{n}\sum_{i=1}^{n} L_i(t)$$

- Regresión:

$$L_i(t) = \{Y_i - \widehat{r}_t^{(-i)}(\mathbf{X}_i)\}^2$$

$$t_{opt} = \arg\min_t \ \mathsf{CV}(t)$$

Cross Validation: K-fold - Representación esquemática (ISLR)



Cross Validation: K folders - Fórmulas

t: tunning parameter

$$\mathsf{CV}(t) = \frac{1}{K}\sum_{k=1}^{K} L_k(t)$$

- Regresión:

$$L_k(t) = \frac{1}{|\mathcal{T}_k^c|}\sum_{j\in\mathcal{T}_k^c}\{Y_j - \widehat{r}_{t,\mathcal{T}_k}(\mathbf{X}_i)\}^2$$

$$t_{opt} = \arg\min_t \ \mathsf{CV}(t)$$

Siendo más formales....

Predicción sin covariables

- Y variable respuesta.
- Esperanza de Y: $\mu = \mathbb{E}(Y)$
- Esperanza desde la predicción.

$$\mu = \arg\min_a \mathbb{E}\{(Y - a)^2\}.$$

Predicción sin covariables

- Y variable respuesta.
- Esperanza de Y: $\mu = \mathbb{E}(Y)$
- Esperanza desde la predicción.

$$\mu = \arg\min_a \mathbb{E}\{(Y - a)^2\}.$$

$$Y_1, \ldots, Y_n \text{ i.i.d. } Y_i \sim Y \quad \longrightarrow \quad \widehat{\mu} = \overline{Y}$$

Predicción - Error cuadrático

- Y: variable respuesta, $X \in \mathbb{R}$: p covariable, $g(\mathbf{X})$ posible predictor.
- Error error cuadrático medio al predecir con g:

$$\mathbb{E}\left[\{Y - g(\mathbf{X})\}^2\right].$$

- Mejor predictor: $r(\mathbf{X})$ satisfaciendo

$$\mathbb{E}\left[\{Y - r(X)\}^2\right] \leq \mathbb{E}\left[\{Y - g(X)\}^2\right], \quad \forall g : \mathbb{R} \to \mathbb{R}$$

$r(X)$ minimiza el error cuadrático medio de predicción

$$r(x) = \mathbb{E}(Y \mid X = x).$$

...the conditional expectation, also known as the regression function. (Hastie & Tibshirani)

Regresión No Paramétrica

$$r(x) = \mathbb{E}(Y \mid X = x).$$

...the conditional expectation, also known as the regression function. (Hastie & Tibshirani)

- Enfoque Paramétrico:
  $r(x) = g(x, \beta)$

Regresión No Paramétrica

$$r(x) = \mathbb{E}(Y \mid X = x).$$

...the conditional expectation, also known as the regression function. (Hastie & Tibshirani)

- Enfoque Paramétrico:
  $r(x) = g(x, \beta)$ por ej.: $g(x, \beta) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$

Regresión No Paramétrica

$$r(x) = \mathbb{E}(Y \mid X = x).$$

...the conditional expectation, also known as the regression function. (Hastie & Tibshirani)

- Enfoque Paramétrico:
  $r(x) = g(x, \beta)$ por ej.: $g(x, \beta) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$

- Enfoque No Paramétrico:
  $r(x)$ no hacemos supuesto de forma, estimación directa

Regresión No Paramétrica

$$r(x) = \mathbb{E}(Y \mid X = x).$$

...the conditional expectation, also known as the regression function. (Hastie & Tibshirani)

- Enfoque Paramétrico:
  $r(x) = g(x, \beta)$ por ej.: $g(x, \beta) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$
  (parameter–driven)

- Enfoque No Paramétrico:
  $r(x)$ no hacemos supuesto de forma, estimación directa
  (data–driven)