Ciencias de Datos con R: Fundamentos Estadísticos

Ana M. Bianco, Jemina García y Mariela Sued.

Regresión no paramétrica

La dulce espera

¿Cuánto medirá al ser adulte?

Regresión - Predicción
Largada: Probabilidades - (sin datos)

Predicción sin variables explicativas (error cuadrático)

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

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

Predicción - Error cuadrático

- Y: variable respuesta, $\mathbf{X}$: variables explicativas, $g(\mathbf{X})$ 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

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

$r(\mathbf{X})$ minimiza el error cuadrático medio de predicción

Predicción - Error cuadrático

- Y: variable respuesta, $\mathbf{X}$: variables explicativas, $g(\mathbf{X})$ 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

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

$r(\mathbf{X})$ minimiza el error cuadrático medio de predicción

¿Quién es $r(\mathbf{X})$?

Esperanza condicional: Función de regresión

$$(\mathbf{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$$

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

...the conditional expectation, also known as the regression function. (SL sin R)

Esperanza condicional: Función de regresión

$$(\mathbf{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$$

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

...the conditional expectation, also known as the regression function. (SL sin R)

$\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$ es la esperanza de la distribución condicional de $Y \mid \mathbf{X} = \mathbf{x}$

¿Qué era la esperanza Condicional?

$$\mathbf{X} \in \mathbb{R}^p \,, r : \mathbb{R}^p \to \mathbb{R}$$

$$Y := r(\mathbf{X}) + \varepsilon \,, \mathbf{X} \text{ independiente de } \varepsilon \,, \mathbb{E}(\varepsilon) = 0.$$

$$r(\mathbf{x}) \quad \text{función de regresión}$$

$$(\mathbf{X}, Y) \in \mathbb{R}^p \times \mathbb{R} \quad \text{vector aleatorio}$$

Función de regresion $r(\mathbf{X})$ - A la carta

$$(\mathbf{X}, Y) \sim \mathcal{P}\ , \quad r(\mathbf{X}) = \mathbb{E}(Y \mid \mathbf{X})$$

$Y := r(\mathbf{X}) + \varepsilon$, $\mathbf{X}$ independiente de ε , $\mathbb{E}(\varepsilon) = 0$.

Función de regresion $r(\mathbf{X})$ - A la carta

$$(\mathbf{X}, Y) \sim \mathcal{P} \;, \quad r(\mathbf{X}) = \mathbb{E}(Y \mid \mathbf{X})$$

$Y := r(\mathbf{X}) + \varepsilon$, $\mathbf{X}$ independiente de ε , $\mathbb{E}(\varepsilon) = 0$.

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

Predecimos con $r(\mathbf{X})$.

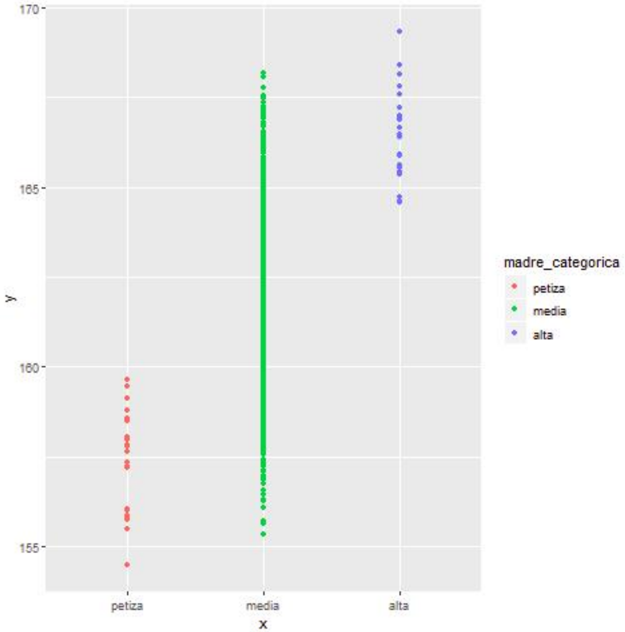
Estadística: Estimación de $r(\mathbf{X})$

$$(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n) \quad \text{iid }, (\mathbf{X}_i, Y_i) \sim P$$

$$\widehat{r}(\,\cdot\,) = \widehat{r}_n(\,\cdot\,) \quad \text{construído con } \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$$

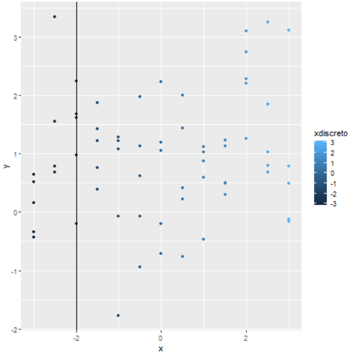
Predecimos con $\widehat{r}_n(\mathbf{X})$.

X discreta


170
165
160
155
y
petiza
media
alta
x
madre_categorica
petiza
media
alta

Estimación no paramétrica de la regresión

- Función de regresión: $r(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$.
- $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, n\}$
- Estimación de $r(\mathbf{x})$ - X discreta.



Estimación no paramétrica de la regresión

- Función de regresión: $r(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$.
- $\{(X_i, Y_i) : i = 1, \ldots, n\}$
- Estimación de $r(\mathbf{x})$ - X discreta.

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

Estimación no paramétrica de la regresión

- Función de regresión: $r(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$.
- $\{(X_i, Y_i) : i = 1, \ldots, n\}$
- Estimación de $r(\mathbf{x})$ - X discreta.

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

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

Estimación no paramétrica de la regresión

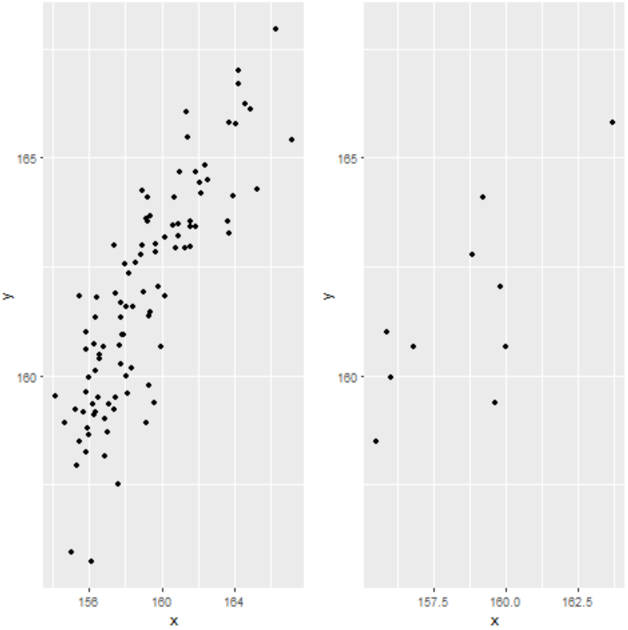
- Función de regresión: $r(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$.
- $\{(X_i, Y_i) : i = 1, \dots, n\}$
- Estimación de $r(\mathbf{x})$ - X discreta.

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

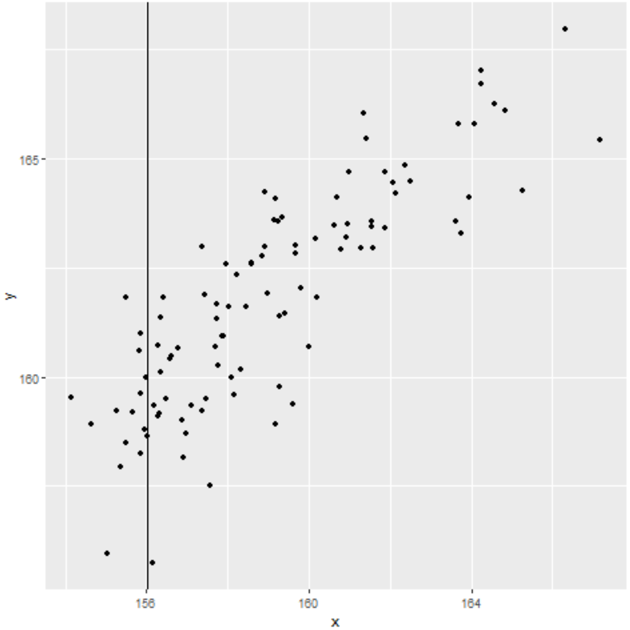
$$\widehat{r}_n(x) = \frac{\sum_{i=1}^{n} Y_i I_{\{|X_i - x| = 0\}}}{\sum_{i=1}^{n} I_{\{|X_i - x| = 0\}}}$$

- X continua?

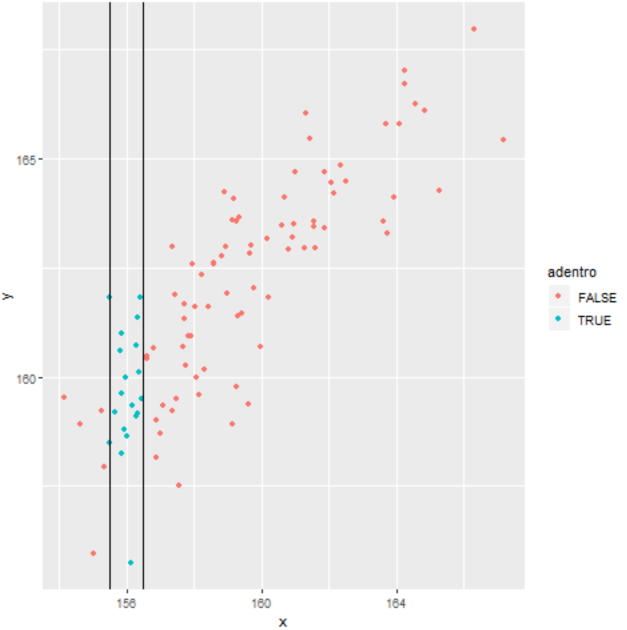
Posibles Escenarios



X continuos: Nadaraya (1964)-Watson(1964)



X continuos: Nadaraya (1964)-Watson(1964)



Nadaraya - Watson kernel regression:

- Estimation of $r(x) = \mathbb{E}(Y \mid X = x)$ - X continuos.

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

Nadaraya - Watson kernel regression:

- Estimation of $r(x) = \mathbb{E}(Y \mid X = x)$ - X continuos.

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

Nadaraya - Watson kernel regression:

- Estimation of $r(x) = \mathbb{E}(Y \mid X = x)$ - X continuos.

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

Nadaraya - Watson kernel regression

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

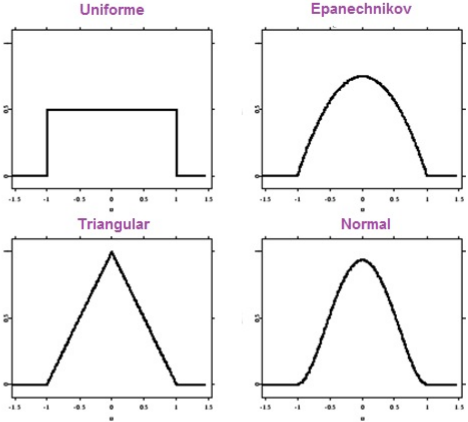
$$\frac{\displaystyle\sum_{i=1}^n Y_i \; K\left(\frac{x - X_i}{h}\right)}{\displaystyle\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)} \quad = \quad \sum_{i=1}^n Y_i \; \underbrace{\frac{K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)}}_{= \; W_i}$$

$$\widehat{r}_n(x) \quad = \quad \sum_{i=1}^n Y_i \, W_i(x)$$ es una media ponderada.

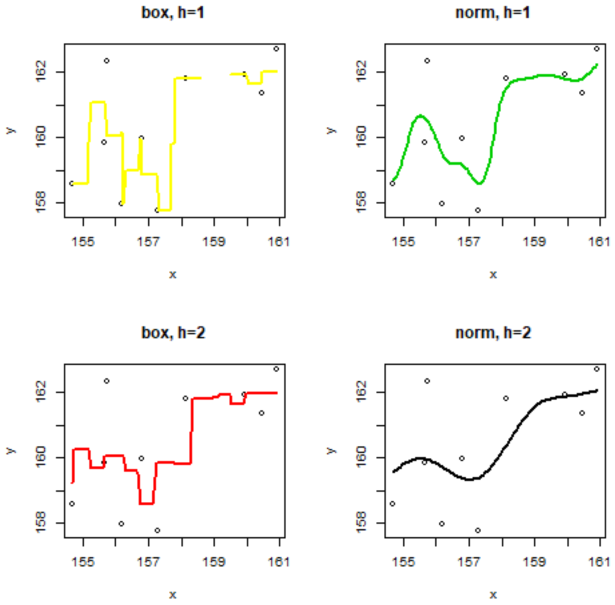
Tipos de núcleos

- 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: $K(t) = \frac{3}{4}(1 - t^2)\mathcal{I}_{[-1,1]}(t)$

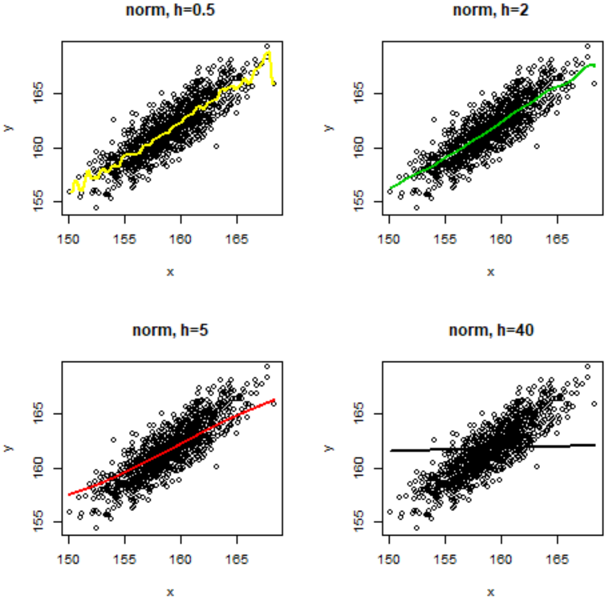
Núcleos



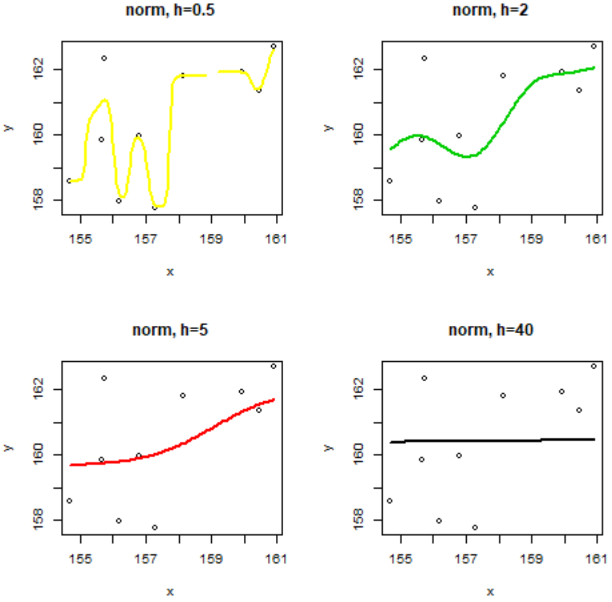
Efecto Nucleo - Pocas



Efecto Ventana - Muchos Datos



Efecto Ventana - Pocos Datos.



Nadaraya - Watson

$$\widehat{r}(\mathbf{x}) = \sum_{i=1}^{n} Y_i \, \frac{K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)}$$

Teorema: $\mathbf{X} \in \mathbb{R}^p$. Si $n \to \infty$, $h \to 0$ y $nh^p \to \infty$, entonces

$$\widehat{r}(\mathbf{x}) \to r(\mathbf{x}) \; , \quad r(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}) \; ,$$

under mild regularity conditions on the joint probability distribution Pr$(\mathbf{X}, Y)$.

otra opción: $Y = r(\mathbf{X}) + \varepsilon$

Tuning Parameter: ventana h

En la práctica, ¿cómo elegimos la ventana?

knn- Vecinos más cercanos - Stone (1977)

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

knn- Vecinos más cercanos - Stone (1977)

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

- 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.

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

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

k-nn: Aproximador Universal

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

Teorema: As $N, k \to \infty$ such that $k/N \to 0$,

$$\widehat{r}_k(\mathbf{x}) \longrightarrow r(\mathbf{x}) \ , \quad r(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}) \ ,$$

under mild regularity conditions on the joint probability distribution Pr$(\mathbf{X}, Y)$.

- No free lunch Theorem (Devroy 1982): No uniform Rates.
- Rates under different kind of assumptions. Estadística Matemática.

Tuning Parameter: cantidad de vecinos k

En la práctica, ¿cómo elegimos la cantidad de vecinos?

Tunning parameter

Siempre nos faltan dos mangos para el peso

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

Tunning parameter selection

Todo muy lindo, pero ...

¿Qué hago con mis datos?

Test Error vs. Training Error

- $\mathcal{D}_n\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$, $\widehat{r} = \widehat{r}_{t,\mathcal{D}_n} = \widehat{r}_n$.
  Predicción:$\widehat{r}_{t,\mathcal{D}_n}(\mathbf{X})$
- Test error:

$$\mathbb{E}\left(\{Y_{\textsf{new}} - \widehat{r}_{t,\mathcal{D}_n}(\mathbf{X}_{\textsf{new}})\}^2 \mid \mathcal{D}_n\right\}$$

Test Error vs. Training Error

- $\mathcal{D}_n\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$, $\widehat{r} = \widehat{r}_{t,\mathcal{D}_n} = \widehat{r}_n$.
  Predicción:$\widehat{r}_{t,\mathcal{D}_n}(\mathbf{X})$
- Test error:

$$\mathbb{E}\left(\{Y_{\text{new}} - \widehat{r}_{t,\mathcal{D}_n}(\mathbf{X}_{\text{new}})\}^2 \mid \mathcal{D}_n\right)$$

- $i-$ésimo Error Cuadrático de Predicción:

$$\{Y_i - \widehat{r}_{t,\mathcal{D}_n}(\mathbf{X}_i)\}^2$$

- Training Error - Error Cuadrático de Predicción Promediado

$$\text{ECPP(t)} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \widehat{r}_{t,\mathcal{D}_n}(\mathbf{X}_i)\}^2$$

Muestra de entrenamiento: observaciones que se utilizan para construir $\widehat{r}$.

Splitting the data

do as well as possible within a given class of rules... This is achieved by splitting the data into a training sequence and a testing sequence. (DGL)

Data-rich situation

- The training set is used to fit the models
- The validation set is used for model selection
- The test set is used for assessment of the generalization error of the final chosen model.

Data-rich situation- Tunning Parameter Selection

- $\mathcal{T}$: the training set is used to fit the models. 80%:
- $\mathcal{V}$: the validation set is used for model selection. 20%

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

$$t_{\mathsf{opt}} = \operatorname{argmin} \widehat{L(t)}$$

¿ Cómo elegimos el tamaño de la ventana?

Validación cruzada 80% -20%

Alt madre (cm)	Alt hijO (cm)
...	...
155.8	170.9
155.8	173.5
157.1	170.7
157.7	171.5
158.1	176.3
158.3	168.3
158.7	171.1
159.7	172
159.7	175.1
159.8	172.7
159.8	174.1
160.5	169.3
160.8	170.5
160.9	177.2
161.2	170.7
162.3	174.6
162.4	173.3
162.5	177.1
163.2	174.5
163.6	173.7
...	...

¿ Cómo elegimos el tamaño h de la ventana?

Training 80% - Validation 20%

Alt madre (cm)	Alt hijO (cm)
...	...
155.8	170.9
155.8	173.5
157.1	170.7
157.7	171.5
158.1	176.3
158.3	168.3
158.7	171.1
159.7	172
159.7	175.1
159.8	172.7
159.8	174.1
160.5	169.3
160.8	170.5
160.9	177.2
161.2	170.7
162.3	174.6
162.4	173.3
162.5	177.1
163.2	174.5
163.6	173.7
...	...

Validación Cruzada - $h = 1$

Predigo en las madres rojas utilizando SOLO los datos negros y $h = 1$

Alt madre(cm)	Predicción con h=1	Alt hijo OBSERVADA
158.7	172.383	171.1
160.5	172.7	169.3
162.3	174.64	174.6
163.2	174.64	174.5

Validación Cruzada - $h = 1$

Predigo en las madres rojas utilizando SOLO los datos negros y $h = 1$

Alt madre(cm)	Predicción con h=1	Alt hijo OBSERVADA
158.7	172.383	171.1
160.5	172.7	169.3
162.3	174.64	174.6
163.2	174.64	174.5

Error(con h=1): =sumo $\{$predichos (con h=1) - observados$\}^2$.

Error(con h=1): $=(172.383 - 171.1)^2 + (172.7 - 169.3)^2 + (174.64 - 174.6)^2 + (174.64 - 174.5)^2 =$ 13.22729

Validación Cruzada - $h = 1.5$

Predigo en las madres rojas utilizando SOLO los datos negros y $h = 1.5$

Alt madre(cm)	Predicción con h=1.5	Alt hijo OBSERVADA
158.7	171.58	171.1
160.5	172.7	169.3
162.3	173.95	174.6
163.2	174.64	1174.5

Error(con h=1.5): =sumo {predichos (con h=1.5) - observados}^{2}.

Error(con h=1.5): $=(171.58 - 171.12)^2 + (172.7 - 169.3)^2 + (173.95 - 174.6)^2 + (174.64 - 174.5)^2 =$ 12.0021

Selección de ventana por Validación cruzada: $h = 1$ o $h = 1.5$?

Error(con h=1.5): =12.0021

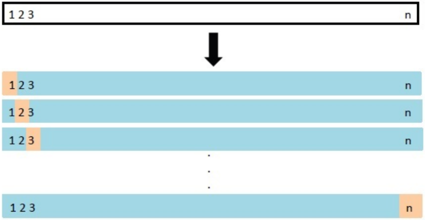
Error(con h=1): =13.22729

Y el ganador es: h=1.5 !!!!!

Menos Datos- Tunning Parameter Selection

Cross Validation

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



Cross Validation: Leave one out - Fórmulas

t: tunning parameter

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

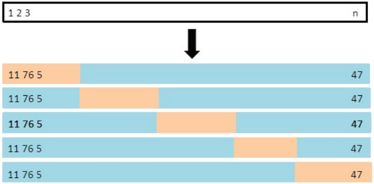
cuidado con n

- Regresión:

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

$$t_{\mathsf{opt}} = \operatorname{argmin} \widehat{L(t)}$$

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



Cross Validation: K folders - Fórmulas

t: tunning parameter

$$\widehat{L(t)} = \frac{1}{K}\sum_{k=1}^{K} L_k(t)$$

- Regresión:

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

$$t_{\textsf{opt}} = \operatorname{argmin} \widehat{L(t)}$$

Para la próxima

- No entendí, ¿Cómo elegimos el tamaño h de la ventana para hacer la barrita?
- ¿Que hacemos con la altura del papá?
- ¿Cómo incluímos otras variables?
- ¿Qué onda con **cuandrados mínimos**?
- ¿Y los modelos?

La altura del canario

Atención: No siempre más es mejor. Hay información que puede ser no relevante. No ayuda a predecir mejor.

¿Qué hicimos para elegir el tamaño h de la ventana?

1. Conjunto de entrenamiento: seleccione al azar el 80% de las filas de su base de datos.
2. Para cada una de las madres no seleccionadas, realice una predicción de la altura del hijo utilizando el conjunto de entrenamiento.
3. Calcule la diferencia entre el valor predicho y el dato de la tabla de la altura del hijo para cada una de las madres no seleccionadas.
4. Considere el error dado por la suma de los cuadrados de las diferencias calculadas en al item anterior.
5. Determine para que tamaño de ventana obtiene el error más chico.