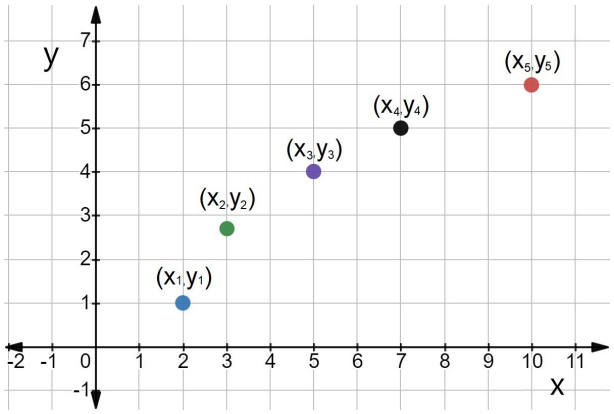
Capítulo 4

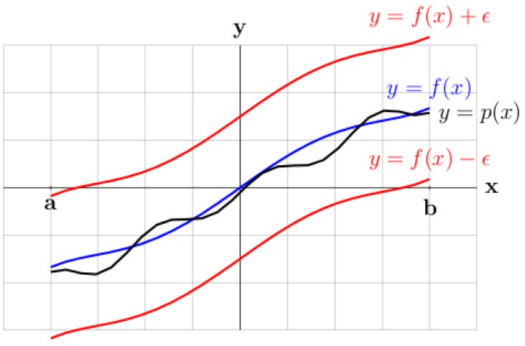
Interpolación

Consideremos la situación en la que nos encontramos ante una secuencia discreta de datos, como la de la siguiente figura



Si no disponemos de la función que dio origen a estos datos o, simplemente, esta función no existe, necesitaremos elaborar alguna estrategia si queremos estimar qué valor tomaría "y" para algún valor intermedio de las abscisas. Una posibilidad (hay otras) sería obtener una función continua que pase por los puntos que ya conocemos. Los polinomios son una buena opción para ello. Dentro de las razones por las cuales los polinomios resultan particularmente útiles tenemos:

1. Teorema de aproximación de Weierstrass: Sea $f : \mathbb{R} \to \mathbb{R}$ continua en $[a, b]$, para cada $\varepsilon > 0$, existe un polinomio $p(x)$ tal que
$$|p(x) - f(x)| < \varepsilon, \quad \forall x \in [a, b]. \tag{4.1}$$



2. Son fáciles de derivar e integrar, y sus derivadas e integrales son a su vez polinomios.

Un candidato de función continua a utilizar podría ser el polinomio de Taylor pero un inconveniente importante que presenta es que concentra toda la información alrededor del punto a partir del cual se realiza el desarrollo (además de requerir el conocimiento de la función y que ésta sea derivable), con lo que puede dar malos resultados para valores de abscisa alejados de ese punto.

A continuación veremos cómo encontrar mejores polinomios que los de Taylor para aproximar a la función subyacente en puntos en los que no disponemos de datos.

4.1. Polinomios de Lagrange

Consideremos la construcción del polinomio $P(x)$ de grado máximo n que pasa por los $n+1$ puntos

$$(x_0, f(x_0)),\ (x_1, f(x_1)), \ldots, (x_n, f(x_n)), \tag{4.2}$$

donde los *nodos* x_i, $i = 0, 1, ..., n$, son todos distintos entre sí. A ese polinomio lo podemos definir como

$$\boxed{P(x) = \sum_{k=0}^{n} f(x_k) L_k(x),} \tag{4.3}$$

con

$$L_k(x) = \frac{(x-x_0)(x-x_1)...(x-x_{k-1})(x-x_{k+1})...(x-x_n)}{(x_k-x_0)(x_k-x_1)...(x_k-x_{k-1})(x_k-x_{k+1})...(x_k-x_n)} = \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{(x-x_i)}{(x_k-x_i)}, \tag{4.4}$$

y recibe el nombre de **n-ésimo polinomio de Lagrange**[1].

Observaciones

• Los $L_k(x)$, $k = 0, 1, ..., n$, son polinomios de grado n, por lo que es evidente que $P(x)$ será un polinomio de grado máximo igual a n.

• Es posible ver que por $n+1$ puntos pasa un único polinomio de grado n o menor. Supongamos que existieran 2 polinomios de grado n que pasaran por $n+1$ puntos. Entonces, la diferencia también sería un polinomio de grado n y ese polinomio tendría $n+1$ ceros, lo cual violaría el Teorema Fundamental del Álgebra.

Ejemplo 1

Construir el polinomio de Lagrange $P(x)$ que pasa por los puntos

$$(x_0, y_0) = (1, 1), \qquad \text{y} \qquad (x_1, y_1) = (4, 7). \tag{4.5}$$

Comenzamos por armar los polinomios $L_k(x)$:

$$L_0 = \frac{x-x_1}{x_0-x_1} = \boxed{\frac{x-4}{1-4}}, \qquad\qquad L_1 = \frac{x-x_0}{x_1-x_0} = \boxed{\frac{x-1}{4-1}}, \tag{4.6}$$

con los cuales construimos el polinomio de Lagrange

$$\begin{aligned} P(x) &= f(x_0)L_0(x) + f(x_1)L_1(x) \\ &= 1 \cdot \frac{x-4}{1-4} + 7 \cdot \frac{x-1}{4-1} \\ &= \left(-\frac{1}{3} + \frac{7}{3}\right) x + \frac{4}{3} - \frac{7}{3} \\ &= \boxed{2x - 1.} \end{aligned} \tag{4.7}$$

[1] Analizaremos en clase cómo los polinomios de Lagrange alcanzan el objetivo propuesto.

Ejemplo 2

Supongamos que queremos aproximar a la función $f(x) = 1/x$ a través de un polinomio de Lagrange que intersecte a la misma en $x_0 = 2$, $x_1 = 2{,}5$ y $x_2 = 4$, a fin de estimar su valor en $x = 3$. Para ello, nuevamente debemos calcular primero los $L_k(x)$:

$$L_0(x) = \frac{(x-2{,}5)(x-4)}{(2-2{,}5)(2-4)}, \qquad L_1(x) = \frac{(x-2)(x-4)}{(2{,}5-2)(2{,}5-4)}, \qquad L_2(x) = \frac{(x-2)(x-2{,}5)}{(4-2)(4-2{,}5)}, \tag{4.8}$$

con los que construimos el polinomio de Lagrange

$$P(x) = \sum_{k=0}^{2} f(x_k) L_k(x). \tag{4.9}$$

Haciendo las cuentas encontramos que

$$P(3) = 0{,}325 \tag{4.10}$$

* * *

¿Qué error se comete al aproximar de esta manera?

Teorema (S/D). Sean $\{x_0, x_1, ..., x_n\}$, $n+1$ números distintos en $[a, b]$ y sea $f \in C^{n+1}[a, b]$. Entonces, para cada $x \in [a, b]$ existe un $\alpha(x) \in (a, b)$ (generalmente desconocido) tal que

$$f(x) = P(x) + \frac{f^{(n+1)}(\alpha(x))}{(n+1)!}(x - x_0)(x - x_1)...(x - x_n), \tag{4.11}$$

donde

$$P(x) = \sum_{k=0}^{n} f(x_k) \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{(x - x_i)}{(x_k - x_i)}. \tag{4.12}$$

• Observen que la fórmula del error es similar a la de Taylor, con la diferencia que en el último caso la información está centrada en x_0 (por eso aparece $(x - x_0)^{n+1}$), mientras que en los polinomios de Lagrange utilizan la información en los puntos $x_0, x_1, ..., x_n$.

• Un inconveniente que se tiene para usar esta fórmula es que generalmente NO se conoce la forma de $f^{(n+1)}(x)$, por lo que es difícil acotar el error. La importancia de la fórmula (del error de interpolación) reside en su utilización, por ejemplo, para el cálculo de errores en la integración numérica, lo que veremos más adelante.

• La forma práctica de aproximar un valor es calcular distintas aproximaciones generando polinomios interpoladores que involucren cada vez más datos, hasta lograr una estimación aceptable (que coincida determinado número de decimales en el resultado).

Ejemplo

Veamos cómo se trabaja si no se conoce $f(x)$ y, por ende, tampoco sus derivadas. Supongamos que tenemos una tabla de valores de la función en cinco puntos.

i	x_i	$f(x_i)$
0	1.0	0.7651977
1	1.3	0.6200860
2	1.6	0.4554022
3	1.9	0.2818186
4	2.2	0.1103623

Queremos calcular $f(1{,}5)$. Como no tenemos $f^{n+1}(x)$ es imposible usar la fórmula del error, pero lo que sí se puede hacer es ir construyendo polinomios de Lagrange de grado cada vez mayor e ir comparando el resultados con los de grado menor. Comencemos por construir un polinomio de grado 1, para lo cual siempre conviene elegir los 2 datos más próximos al que nos interesa calcular. En nuestro caso, los datos más próximos a $x = 1{,}5$ serán $x_1 = 1{,}3$ y $x_2 = 1{,}6$, con lo cual armamos el polinomio

$$P_1(x) = \frac{(x-1{,}6)}{(1{,}3-1{,}6)} f(1{,}3) + \frac{(x-1{,}3)}{(1{,}6-1{,}3)} f(1{,}6) \tag{4.13}$$

tal que

$$P_1(1{,}5) = 0{,}5102968. \tag{4.14}$$

Para construir polinomios de grado 2, tenemos dos alternativas. Por un lado, utilizar los puntos $\{x_0, x_1, x_2\}$:

$$P_2(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}f(x_2), \tag{4.15}$$

o bien, utilizar $\{x_1, x_2, x_3\}$:

$$P_2'(x) = \frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)}f(x_1) + \frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)}f(x_2) + \frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)}f(x_3). \tag{4.16}$$

Con estos polinomios, obtenemos

$$P_2(1{,}5) = 0{,}5124715, \tag{4.17}$$

$$P_2'(1{,}5) = 0{,}5112857. \tag{4.18}$$

Repitiendo el proceso, pero ahora para 4 puntos:

$$\{x_0, x_1, x_2, x_3\} \quad \longrightarrow \quad P_3(1{,}5) = 0{,}5118127, \tag{4.19}$$

$$\{x_2, x_2, x_3, x_4\} \quad \longrightarrow \quad P_3'(1{,}5) = 0{,}5118302, \tag{4.20}$$

y, finalmente, con todos los puntos

$$\{x_0, x_1, x_2, x_3, x_4\} \quad \longrightarrow \quad P_4(1{,}5) = 0{,}5118200. \tag{4.21}$$

A medida que vamos generando polinomios podemos ir comparando los valores obtenidos para la abscisa que nos interesa. Así, vemos cómo cada vez hay más decimales que no cambian y que, por lo tanto, parecen "correctos". A priori uno tendería a pensar que $P_4(x)$ debería ser el polinomio que nos ofrece el resultado más preciso, sin embargo, esto no necesariamente es cierto. La función tabulada es la función de Bessel de primer tipo y orden cero, cuyo valor para $x = 1{,}5$ es 0,5118277. Se ve que el resultado más próximo al real fue el de P_3'.

* * *

Como vimos en el ejemplo anterior, hasta ahora nunca utilizamos la información del polinomio de grado k para obtener el de grado $k+1$, ni la de éste para obtener el de grado $k+2$. Esto implica que cada vez que uno desea incorporar un nuevo punto para construir un polinomio de Lagrange, tiene que hacer todas las cuentas desde cero. A continuación mostraremos que es posible basarnos en polinomios de Lagrange de grado menor para construir uno de grado mayor. Antes de hacerlo, necesitamos definir una nueva notación.

Definición. Sea $f(x)$ una función definida en los $n+1$ puntos $x_0, x_1, ..., x_n$ y sean m_i, $i = 1, ..., k$, k enteros distintos que verifican que $0 \leq m_i \leq n$, $\forall i$. Denotaremos por

$$P_{m_1, m_2, ..., m_k}(x), \tag{4.22}$$

al polinomio de Lagrange que coincide con $f(x)$ en los k puntos $x_{m_1}, x_{m_2}, ..., x_{m_k}$.

Ejemplo

En virtud de la notación introducida, el polinomio de Lagrange $P_{1,2,4}(x)$ que coincide con $f(x)$ en los puntos x_1, x_2 y x_4 será

$$P_{1,2,4}(x) = f(x_1)\frac{(x-x_2)(x-x_4)}{(x_1-x_2)(x_1-x_4)} + f(x_2)\frac{(x-x_1)(x-x_4)}{(x_2-x_1)(x_2-x_4)} + f(x_4)\frac{(x-x_1)(x-x_2)}{(x_4-x_1)(x_4-x_2)}. \tag{4.23}$$

Podemos presentar entonces el siguiente teorema.

Teorema. Sea $f(x)$ una función definida en $\{x_0, x_1, ..., x_n\}$ y sean x_i y x_j, dos números distintos dentro del conjunto. Entonces, el n-ésimo polinomio de Lagrange que interpola a $f(x)$ entre los $n+1$ puntos $\{x_0, x_1, ..., x_n\}$ está dado por

$$P(x) = \frac{(x - x_j)P_{0,1,2,...,j-1,j+1,...,n}(x) - (x - x_i)P_{0,1,2,...,i-1,i+1,...,n}(x)}{x_i - x_j}. \tag{4.24}$$

Demostración

Para simplificar un poco la notación, vamos a hacer

$$Q^i(x) \equiv P_{0,1,2,...,i-1,i+1,...,n}(x), \tag{4.25}$$

$$Q^j(x) \equiv P_{0,1,2,...,j-1,j+1,...,n}(x). \tag{4.26}$$

Como $Q^i(x)$ y $Q^j(x)$ son polinomios de grado $n-1$ (o menor, pero nunca mayor), $P(x)$ tendrá grado máximo n. Consideremos $0 \leq r \leq n$ tal que $r \neq i$ y $r \neq j$. Por construcción, sabemos que

$$Q^i(x_r) = Q^j(x_r) = f(x_r), \tag{4.27}$$

luego

$$P(x_r) = \frac{(x_r - x_j)Q^j(x_r) - (x_r - x_i)Q^i(x_r)}{x_i - x_j} = \frac{x_i - x_j}{x_i - x_j} f(x_r) = f(x_r). \tag{4.28}$$

Además, tenemos que $Q^j(x_i) = f(x_i)$, luego

$$P(x_i) = \frac{(x_i - x_j)Q^j(x_i) - \cancel{(x_i - x_i)Q^i(x_i)}}{x_i - x_j} = f(x_i). \tag{4.29}$$

Análogamente, como $Q^i(x_j) = f(x_j)$, resulta que $P(x_j) = f(x_j)$. Ahora, por definición, $P_{0,1,2,...,n}(x)$ es el único polinomio de grado n que coincide con $f(x)$ en $x_0, x_1, ..., x_n$, por lo tanto

$$\boxed{P(x) = P_{0,1,2,...,n}(x).} \tag{4.30}$$

* * *

Como consecuencia de este teorema, tenemos una forma recursiva (es decir, utilizando información previa) de construir los polinomios de Lagrange. Este mecanismo se llama método de Neville.

Método de Neville

Veamos cómo proceder en la práctica. Supongamos que conocemos 5 puntos de la función: $(x_0, f(x_0))$, $(x_1, f(x_1))$, $(x_2, f(x_2))$, $(x_3, f(x_3))$ y $(x_4, f(x_4))$. Los sucesivos polinomios de Lagrange pueden ir construyéndose con una tabla como la de abajo, completando líneas de arriba hacia abajo para finalmente obtener el polinomio que comprende a todos los puntos.

x_0	P_0				
x_1	P_1	$P_{0,1}$			
x_2	P_2	$P_{1,2}$	$P_{0,1,2}$		
x_3	P_3	$P_{2,3}$	$P_{1,2,3}$	$P_{0,1,2,3}$	
x_4	P_4	$P_{3,4}$	$P_{2,3,4}$	$P_{1,2,3,4}$	$P_{0,1,2,3,4}$

* * *

Como dijimos antes, a menudo se cree que la utilización de un polinomio de mayor grado asegura una mejor aproximación. Esto en general no es así. Con los polinomiosde Lagrange, a medida que uno aumenta el número de puntos también aumenta el grado y es aquí donde la naturaleza oscilatoria de los polinomios nos puede jugar una mala pasada (*efecto Runge* o *fenómeno Runge*).

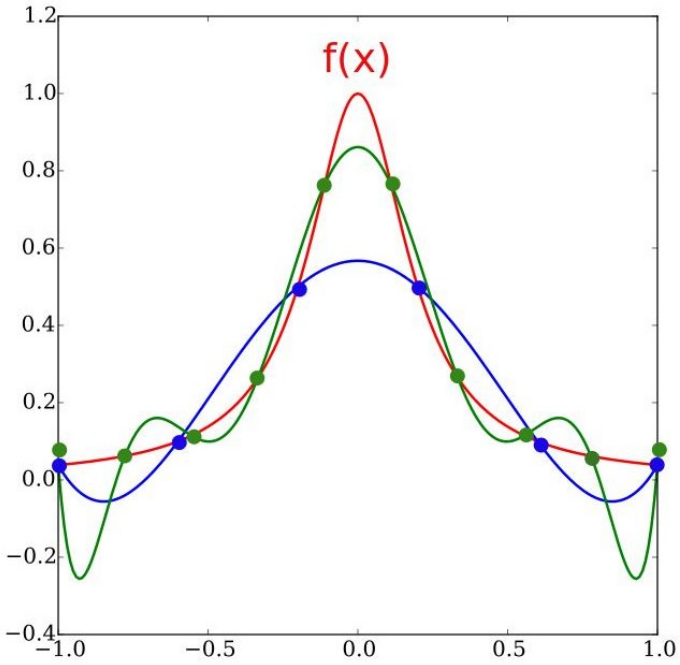


Figura 4.1: Efecto Runge.

Para evitar este problema uno puede recurrir a la estrategia que explicamos a continuación.

4.2. Interpolación cúbica segmentaria (splines)

Introducción (en clase)

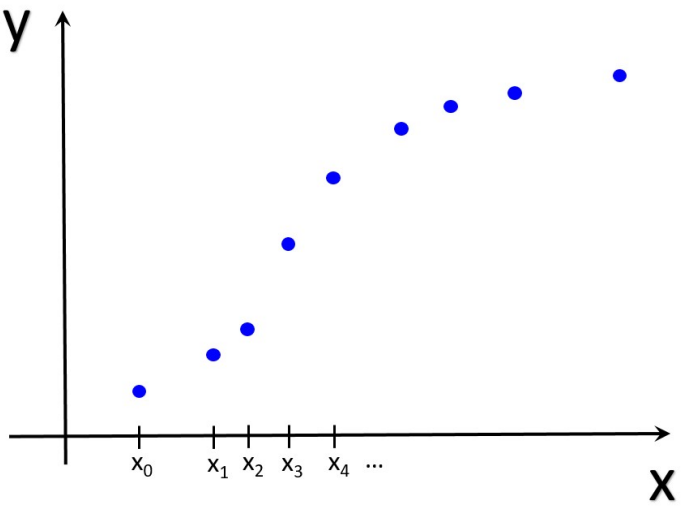


Figura 4.2: Interpolación cúbica segmentaria

Definición. Dada una función $f(x)$ definida en $[a,b]$ y un conjunto de $n+1$ nodos

$$a = x_0 < x_1 < x_2 < ... < x_n = b, \tag{4.31}$$

un interpolador polinómico cúbico segmentario $S(x)$ para $f(x)$ es una función que satisface las siguientes condiciones:

(a) $S(x)$ es un polinomio cúbico, denotado $S_j(x)$ en el subintervalo $[x_j, x_{j+1}]$, donde $j = 0, 1, 2, ..., n-1$

(b) $S_j(x_j) = f(x_j)$ y $S_j(x_{j+1}) = f(x_{j+1})$ para $j = 0, 1, 2, ..., n-1$

(c) $S_j(x_{j+1}) = S_{j+1}(x_{j+1})$ para $j = 0, 1, 2, ..., n-2$

(d) $S'_j(x_{j+1}) = S'_{j+1}(x_{j+1})$ para $j = 0, 1, 2, ..., n-2$

(e) $S''_j(x_{j+1}) = S''_{j+1}(x_{j+1})$ para $j = 0, 1, 2, ..., n-2$

(f) una de las siguientes condiciones de borde es satisfecha

(I) $S''_j(x_0) = S''_j(x_n)$ (condición natural o libre)

(II) $S'_j(x_0) = f'(x_0)$ y $S'_j(x_n) = f'(x_n)$ (condición sujeta)

• Es posible probar que los coeficientes de todos los polinomios interpoladores, que tienen la forma

$$S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3, \tag{4.32}$$

pueden hallarse resolviendo un **sistema lineal** de $(n+1)\times(n+1)$ **tridiagonal** (donde $n+1$ es el número de datos), que siempre es **diagonal dominante**. Consecuentemente, se puede demostrar (usando el teorema de Gerschgorin, que veremos más adelante) que el sistema tiene solución y ésta es única.

4.3. Cuadrados mínimos

Consideremos la situación que se observa en los siguientes gráficos, donde los puntos y_i, $i = 1, ..., m$, pueden contener errores.

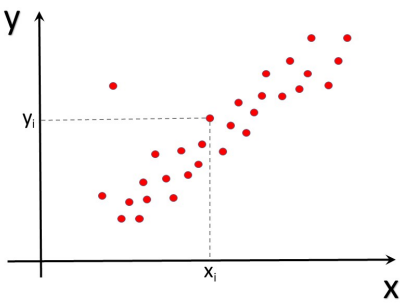


Figura 4.3: Datos A.

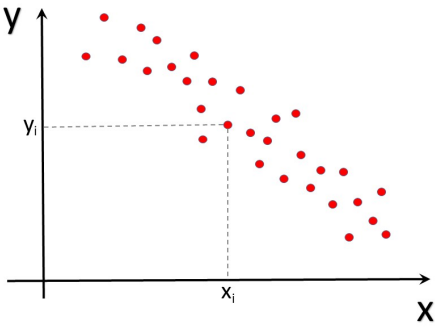


Figura 4.4: Datos B.

En este caso, no tiene sentido interpolar exactamente todos los puntos mediante un polinomio de grado $m-1$, ni intentar realizar un ajuste por *splines*. Noten, sin embargo, que los puntos graficados parecen seguir un comportamiento general de la forma

$$y = a_1 x + a_0, \tag{4.33}$$

lo que de alguna manera refuerza la idea de abandonar la interpolación tal como la vimos. Lo que es evidente es que una recta no va a pasar por todos los puntos, de modo que debemos elegir algún criterio que nos permita determinar con cuál de todas las posibles rectas quedarnos. Existen diferentes criterios que se pueden elegir. A continuación mencionamos algunos.

4.3.1. Criterios para elegir "el mejor ajuste"

1. Podemos intentar hallar el mejor ajuste buscando los valores de a_0 y a_1 que minimicen el estimador de error

$$E_\infty(a_1, a_0) = \max_{\forall i} \{|y_i - (a_1 x_i + a_0)|\}. \tag{4.34}$$

   De esta manera, buscamos la recta cuya distancia al punto de peor ajuste sea la mínima. A este criterio se lo denomina **MINIMAX** y se caracteriza por ser difícil de implementar.

2. Otra estrategia se basa en hallar a_0 y a_1 que minimicen el error total

$$E_1(a_1, a_0) = \sum_{i=1}^{n} |y_i - (a_1 x_i + a_0)|. \tag{4.35}$$

   $E_1(a, b)$ se llama **desviación absoluta**, y para encontrar el mínimo debemos calcular

$$\frac{\partial E_1}{\partial a_1} = 0, \qquad \text{y} \qquad \frac{\partial E_1}{\partial a_0} = 0, \tag{4.36}$$

   lo que implica derivar la función valor absoluto, que en el cero no es derivable, por lo que podemos llegar a tener dificultades para encontrar la solución.

3. Una tercera alternativa es buscar el par de parámetros que minimicen la suma del cuadrado de los errores

$$E_2(a_1, a_0) = \sum_{i=1}^{n} \left[y_i - (a_1 x_i + a_0)\right]^2, \tag{4.37}$$

   que es la llamada **aproximación de cuadrados mínimos**. Esta es la opción con la que vamos a trabajar. La recta $y = a_1 x + a_0$ con la que nos quedaremos será la que minimice la suma de los cuadrados de las distancias entre el "dato experimental", y_i, y el valor predicho por la recta $a_1 x_i + a_0$, para la misma abscisa. Esta opción se elige no sólo porque minimizar $E_2(a_1, a_0)$ es matemáticamente más simple que las opciones anteriores, sino porque estadísticamente es mejor y, además, provee de una distribución más adecuada de los pesos relativos de los errores comparado con los métodos anteriores.

4.3.2. Búsqueda de parámetros en cuadrados mínimos

Para hallar a_0 y a_1 hacemos

$$\frac{\partial E_2}{\partial a_0} = 2\sum_{i=1}^{m}\big[y_i - (a_1x_i + a_0)\big](-1) = 0 \tag{4.38}$$

$$\frac{\partial E_2}{\partial a_1} = 2\sum_{i=1}^{m}\big[y_i - (a_1x_i + a_0)\big](-x_i) = 0 \tag{4.39}$$

y, por lo tanto, habrá que resolver el sistema

$$\begin{cases} a_1\displaystyle\sum_{i=1}^{m} x_i + a_0\sum_{i=1}^{m} 1 = \sum_{i=1}^{m} y_i, \\ a_1\displaystyle\sum_{i=1}^{m} x_i^2 + a_0\sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_ix_i. \end{cases} \tag{4.40}$$

Este sistema puede reescribirse en forma matricial definiendo

$$\boldsymbol{A} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ & \vdots \\ 1 & x_m \end{pmatrix}, \qquad \boldsymbol{b} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \qquad \boldsymbol{x} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}, \tag{4.41}$$

de modo que el sistema de dos ecuaciones con dos incógnitas nos queda simplemente como

$$\boxed{\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b},} \tag{4.42}$$

que es el llamado **sistema de ecuaciones normales**[2] asociado al sistema lineal $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Como las abscisas x_i se suponen distintas entre sí, resulta que el rango de $\boldsymbol{A}$ es igual a 2, lo que permite asegurar que la solución es única (resultado de la teoría de matrices normales) e igual a

$$\boldsymbol{x} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b} = \frac{1}{m\sum_{i=1}^{m} x_i^2 - \left(\sum_{i=1}^{m}\right)^2}\begin{pmatrix} \sum_{i=1}^{m} x_i^2 & -\sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & m \end{pmatrix}\begin{pmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} y_ix_i \end{pmatrix}. \tag{4.43}$$

Noten que

$$\boxed{E_2 = \sum_{i=1}^{n}\big[y_i - (a_1x_i + a_0)\big]^2 = (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) = \sum_{i=1}^{n}\varepsilon_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon},} \tag{4.44}$$

con

$$\varepsilon_i = y_i - (a_1x_i + a_0). \tag{4.45}$$

Pronto veremos que el funcional de cuadrados mínimos tiene la misma forma para otras funciones que poseen distinto número de parámetros. En ese caso, $\boldsymbol{A}$ y $\boldsymbol{b}$ cambian, pero E_2 puede seguir escribiéndose como en esta fórmula.

Antes de ocuparnos de ese caso, presentamos lo que se denomina el "problema general de cuadrados mínimos".

---

[2] Sobre los sistemas de ecuaciones normales:

- Para un sistema de $m \times n$ (m ecuaciones con n incógnitas) $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, el sistema asociado de ecuaciones normales es $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b}$. Este nuevo sistema es de $n \times n$.
- $\boldsymbol{A}^T\boldsymbol{A}x = \boldsymbol{A}^T\boldsymbol{b}$ es siempre compatible, aún cuando $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ no lo sea.
- Cuando $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ es compatible, el conjunto de soluciones coincide con el de $\boldsymbol{A}^T\boldsymbol{A}x = \boldsymbol{A}^T\boldsymbol{b}$.
- $\boldsymbol{A}^T\boldsymbol{A}x = \boldsymbol{A}^T\boldsymbol{b}$ tiene solución única sí y sólo sí $r(\boldsymbol{A}) = n$, en cuyo caso $\boldsymbol{x} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$.
- Cuando $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ es compatible determinado, lo mismo vale para las ecuaciones normales y la única solución a ambos sistemas es $\boldsymbol{x} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$.

Problema general de cuadrados mínimos. Sea $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ y $\boldsymbol{b} \in \mathbb{R}^m$; sea además $\boldsymbol{\varepsilon}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}$, el problema general de cuadrados mínimos consiste en encontrar un vector $\boldsymbol{x}$ que minimice

$$E_2(\boldsymbol{x}) = \sum_{i=1}^{m} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}). \tag{4.46}$$

Cualquier vector que provea un valor mínimo para esta expresión se llama solución de cuadrados mínimos.

Teorema. Bajo las hipótesis del problema general de cuadrados mínimos, se verifica que:

(a) el conjunto de todas las soluciones de cuadrados mínimos es precisamente el conjunto de soluciones de las ecuaciones normales;

(b) la solución de cuadrados mínimos es única sí y sólo sí $r(\boldsymbol{A}) = n$, en cuyo caso está dada por $(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$;

(c) si el sistema $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ es compatible entonces las soluciones del sistema $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ son las mismas que las de cuadrados mínimos.

Demostración

(a) Vamos a demostrar primero que todo vector que sea solución de cuadrados mínimos verifica las ecuaciones normales.

$$\begin{aligned} E_2(\boldsymbol{x}) = \sum_{i=1}^{m} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} &= (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \\ &= (\boldsymbol{A}\boldsymbol{x})^T\boldsymbol{A}\boldsymbol{x} - (\boldsymbol{A}\boldsymbol{x})^T\boldsymbol{b} - \boldsymbol{b}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{b} \\ &= \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{b} - \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{b} \\ &= \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{b}. \end{aligned} \tag{4.47}$$

Para hallar el vector $\boldsymbol{x}$ que minimice la expresión anterior vamos a aplicar las técnicas del análisis matemático para diferenciar a la función matricial

$$f(\boldsymbol{x}) = f(x_1, x_2, ..., x_n) = \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{b}. \tag{4.48}$$

Diferenciar una función matricial es similar a la diferenciación de funciones escalares. Supongamos que tenemos una matriz $\boldsymbol{U} = [u_{ij}]$, entonces

$$\left[\frac{\partial \boldsymbol{U}}{\partial x}\right]_{ij} = \frac{\partial u_{ij}}{\partial x}, \qquad \frac{\partial[\boldsymbol{U} + \boldsymbol{V}]}{\partial x} = \frac{\partial \boldsymbol{U}}{\partial x} + \frac{\partial \boldsymbol{V}}{\partial x}, \qquad \frac{\partial[\boldsymbol{U}\boldsymbol{V}]}{\partial x} = \frac{\partial \boldsymbol{U}}{\partial x}\boldsymbol{V} + \boldsymbol{U}\frac{\partial \boldsymbol{V}}{\partial x}. \tag{4.49}$$

Aplicando estas reglas a la función f

$$\frac{\partial f}{\partial x_i} = \frac{\partial \boldsymbol{x}^T}{\partial x_i}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\frac{\partial \boldsymbol{x}}{\partial x_i} - 2\frac{\partial \boldsymbol{x}^T}{\partial x_i}\boldsymbol{A}^T\boldsymbol{b}, \tag{4.50}$$

y teniendo en cuenta que

$$\frac{\partial \boldsymbol{x}}{\partial x_i} = \boldsymbol{e}_i, \tag{4.51}$$

obtenemos

$$\begin{aligned} \frac{\partial f}{\partial x_i} &= \boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{e}_i - 2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{b} \\ &= 2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{b} \end{aligned}$$

$$= 2(\boldsymbol{A}_{i*}^T \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}_{i*}^T \boldsymbol{b}). \tag{4.52}$$

Como buscamos el mínimo de f, debemos hallar $\boldsymbol{x}$ tal que

$$\frac{\partial f}{\partial x_i} = 0, \qquad \text{para } i = 1, 2, ..., n. \tag{4.53}$$

Por lo tanto, usando notación vectorial, en el mínimo deberá cumplirse que

$$2(\boldsymbol{A}^T \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^T \boldsymbol{b}) = 0 \qquad \Rightarrow \qquad \boxed{\boldsymbol{A}^T \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T \boldsymbol{b}.} \tag{4.54}$$

Consecuentemente, queda demostrado que toda solución de cuadrados mínimos verifica las ecuaciones normales.

Ahora deberemos ver que si $\boldsymbol{x}$ es solución de las ecuaciones normales, entonces minimiza a f. Consideremos un vector $\boldsymbol{z}$, que es solución de las ecuaciones normales

$$\boldsymbol{A}^T \boldsymbol{A}\boldsymbol{z} = \boldsymbol{A}^T \boldsymbol{b}. \tag{4.55}$$

Sea otro vector $\boldsymbol{y} \in \mathbb{R}^n$ que NO es solución de las ecuaciones normales. Definimos al vector diferencia

$$\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{z}, \qquad \Rightarrow \qquad \boldsymbol{y} = \boldsymbol{z} + \boldsymbol{u}, \tag{4.56}$$

y evaluamos

$$\begin{aligned}
f(\boldsymbol{y}) &= f(\boldsymbol{z} + \boldsymbol{u}) \\
&= (\boldsymbol{z} + \boldsymbol{u})^T \boldsymbol{A}^T \boldsymbol{A}(\boldsymbol{z} + \boldsymbol{u}) - 2(\boldsymbol{z} + \boldsymbol{u})^T \boldsymbol{A}^T \boldsymbol{b} + \boldsymbol{b}^T \boldsymbol{b} \\
&= \boxed{\boldsymbol{z}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{z}} + \boldsymbol{z}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{u} + \boldsymbol{u}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{z} + \boldsymbol{u}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{u} - \boxed{2\boldsymbol{z}^T \boldsymbol{A}^T \boldsymbol{b}} - 2\boldsymbol{u}^T \boldsymbol{A}^T \boldsymbol{b} + \boxed{\boldsymbol{b}^T \boldsymbol{b}} \\
&= f(\boldsymbol{z}) + \boldsymbol{z}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{u} + \boldsymbol{u}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{z} + \boldsymbol{u}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{u} - 2\boldsymbol{u}^T \boldsymbol{A}^T \boldsymbol{b} \\
&= f(\boldsymbol{z}) + 2\boldsymbol{u}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{z} + \boldsymbol{u}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{u} - 2\boldsymbol{u}^T \boldsymbol{A}^T \boldsymbol{b} \\
&= f(\boldsymbol{z}) + 2\boldsymbol{u}^T (\cancel{\boldsymbol{A}^T \boldsymbol{A}\boldsymbol{z} - \boldsymbol{A}^T \boldsymbol{b}}) + \boldsymbol{u}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{u} \\
&= f(\boldsymbol{z}) + \boldsymbol{u}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{u} \\
&= f(\boldsymbol{z}) + \boldsymbol{v}^T \boldsymbol{v},
\end{aligned} \tag{4.57}$$

donde hemos definido $\boldsymbol{v} = \boldsymbol{A}\boldsymbol{u}$. Como $\boldsymbol{v} \in \mathbb{R}^m$, $\boldsymbol{v}^T \boldsymbol{v} \geq 0$ y por lo tanto concluimos que

$$f(\boldsymbol{y}) \geq f(\boldsymbol{z}), \quad \forall \boldsymbol{y} \in \mathbb{R}^n \tag{4.58}$$

por lo tanto, concluimos que si un vector es solución de las ecuaciones normales entonces la función f tiene un mínimo en ese vector.

Los puntos (b) y (c) no los demostraremos porque son resultados conocidos de la teoría de ecuaciones normales, que ya hemos enunciado.

4.3.3. Nota sobre implementación en computadoras

El conjunto de soluciones del problema de cuadrados mínimos no suele obtenerse de las ecuaciones normales

$$\boldsymbol{A}^T \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T \boldsymbol{b}, \tag{4.59}$$

ya que generalmente $\boldsymbol{A}^T \boldsymbol{A}$ es muy mal comportada cuando $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ está levemente mal condicionada. Eso se puede ver mejor con un ejemplo:

$$\boldsymbol{A} = \begin{pmatrix} 3 & 6 \\ 1 & 2{,}01 \end{pmatrix}, \qquad \boldsymbol{b} = \begin{pmatrix} 9 \\ 3{,}01 \end{pmatrix}. \tag{4.60}$$

Si se resolviera el sistema con el método de Gauss, con 3 dígitos, obtendríamos $\begin{pmatrix}1\\1\end{pmatrix}$, que coincide con la solución exacta. Sin embargo, si se usa aritmética de 3 dígitos para armar el sistema de ecuaciones normales asociado obtendríamos

$$\begin{pmatrix}10 & 20\\20 & 40\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix} = \begin{pmatrix}30\\60{,}1\end{pmatrix}, \tag{4.61}$$

que resulta ser un sistema incompatible. Por esta razón, las ecuaciones normales son evitadas usualmente en el cálculo numérico, aunque son reconocidas como idea teórica valiosa y dan sustento a ciertos métodos numéricos importantes.

Para encontrar la solución al problema de cuadrados mínimos se pueden utilizar otros métodos, por ejemplo, el de la factorización QR.

4.3.4. Factorización QR para el problema de cuadrados mínimos

En lo siguiente supondremos que $r(\boldsymbol{A}) = n$ y que, por lo tanto, existe una única solución al problema de cuadrados mínimos. Entonces, sabemos que se cumple que $\boldsymbol{A}$ admite factorización QR y que ésta es única

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}. \tag{4.62}$$

Como las columnas de $\boldsymbol{Q}$ forman un conjunto ortonormal

$$\boldsymbol{Q}^T\boldsymbol{Q} = I_n, \tag{4.63}$$

y podemos escribir

$$\boldsymbol{A}^T\boldsymbol{A} = (\boldsymbol{Q}\boldsymbol{R})^T(\boldsymbol{Q}\boldsymbol{R}) = \boldsymbol{R}^T\boldsymbol{R}. \tag{4.64}$$

Entonces, reemplazando en las ecuaciones normales

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b}, \tag{4.65}$$

tenemos

$$\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{x} = \boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{b}. \tag{4.66}$$

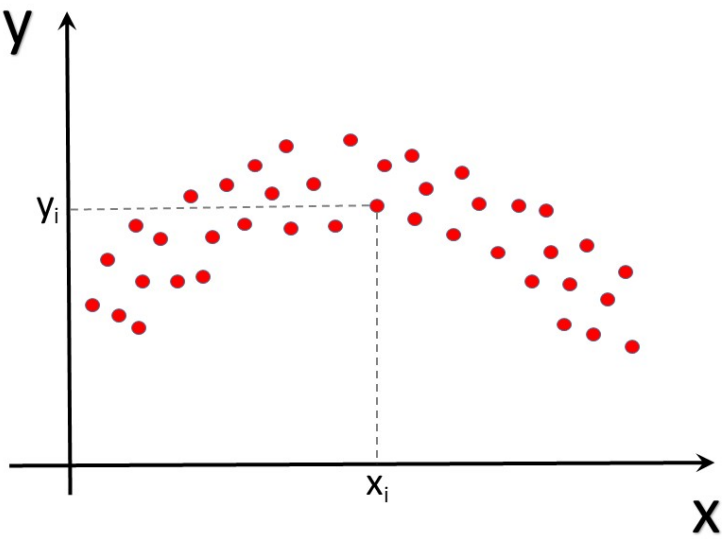
Como $\boldsymbol{R}^T$ es no singular, obtenemos finalmente

$$\boxed{\boldsymbol{R}\boldsymbol{x} = \boldsymbol{Q}^T\boldsymbol{b},} \tag{4.67}$$

que se resuelve por sustitución hacia atrás.

4.3.5. Ajuste no lineal

Consideremos un conjunto de puntos $\{(x_1, y_1), ..., (x_m, y_m)\}$ como el de figura:



Queremos aproximar la "nube de puntos" con el mejor polinomio (en el sentido de cuadrados mínimos) de grado $n-1$, con $n \leq m$. Como en el caso de la recta, supondremos que las abscisas son todas distintas entre sí. Sea

$$p(x) = a_0 + a_1 x + ... + a_{n-1} x^{n-1}, \tag{4.68}$$

de modo que tenemos n parámetos

$$\boldsymbol{\alpha} = \begin{pmatrix} a_0 & a_1 & a_2 & \dots & a_{n-1} \end{pmatrix}^T. \tag{4.69}$$

El error de cuadrados mínimos será

$$E_2(\boldsymbol{\alpha}) = \sum_{i=1}^{m} [p(x_i) - y_i]^2 = (\boldsymbol{A\alpha} - \boldsymbol{y})^T (\boldsymbol{A\alpha} - \boldsymbol{y}), \tag{4.70}$$

donde

$$\boldsymbol{A} = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{n-1} \\ \vdots & \ddots & & & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^{n-1} \end{pmatrix}, \qquad \boldsymbol{\alpha} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{pmatrix} \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}. \tag{4.71}$$

De acuerdo a lo que vimos en el teorema general de cuadrados mínimos, el vector solución $\boldsymbol{\alpha}$ de cuadrados mínimos es obtenido de la solución del sistema de ecuaciones normales asociada al sistema $\boldsymbol{A\alpha} = \boldsymbol{y}$. Como $r(\boldsymbol{A}) = n$ (ya que $\boldsymbol{A}$ es una matriz de Vandermonde con $n \leq m$ y, por lo tanto, sus columnas forman un conjunto linealmente independiente), sabemos que la solución de cuadrados mínimos es única y está dada por

$$\boldsymbol{\alpha} = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{y}. \tag{4.72}$$

* * *

No siempre la función subyacente es lineal con los parámetros. Supongan el caso en el que la aproximación de ajuste requiera ser de la forma

$$y = a_0 \, e^{a_1 x}. \tag{4.73}$$

En este caso conviene trabajar con la forma logarítmica

$$\ln y = \ln a_0 + a_1 x. \tag{4.74}$$

Con esto volvemos a tener un problema lineal, aunque la aproximación que obtenga ya no será la de cuadrados mínimos del problema original.