Método de mínimos cuadrados

Alfredo Villagrán Olguín

Abril, 2020

Interpolación

En el análisis numérico, la interpolación es una forma de estimación; es decir, es un método para pronosticar nuevos datos, dentro de un un conjunto discreto de datos conocidos.

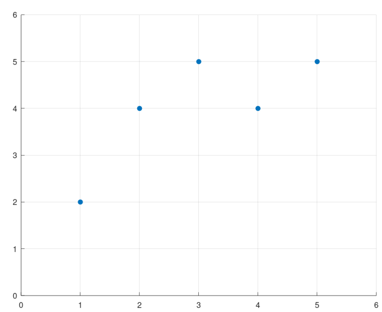


Figura 1

Tanto en la ciencia como en la ingeniería, constantemente se cuenta con un número de datos, cuyo origen pueden ser experimentos, mediciones o muestreos. Al ser datos obtenidos experimentalmente, se desconcoce cuál es la manera en la que se comportan dichas mediciones.

Por lo cual, se requiere calcular una función que pueda acercarse a representar tanto los datos experimentales como pronosticar otros que no se hayan medido pero que se encuentren dentro del rango de estas mediciones.

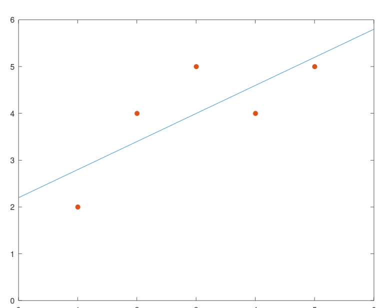


Figura 2

La función calculada es llamada *función de aproximación*. Si la función de aproximación es una recta, a ésta se le llama *aproximación lineal*. Si la función no es una recta, se le llama *aproximación no lineal*. Las aproximaciones no lineales pueden ajustarse a funciones polinomiales, senoidales, exponenciales, etc.

Considerando, por ejemplo, los datos mostrados en la Figura 1, una aproximación lineal a los datos medidos sería la recta que se muestra en la Figura 2. Los métodos de interpolación buscan calcular la función que *mejor* se acerque a todos los puntos. Una vez que se ha calculado la función de la recta, es posible pronosticar para valores no medidos cuál sería su comportamiento, en terminos de la función calculada.

Asimismo, es posible que la cantidad de mediciones sea mayor y que el comportamiento no se ajuste a una línea recta; es decir, que la aproximación no sea lineal. Con los métodos de interpolación, también es posible calcular funciones no lineales que se ajusten al comportamiento de las mediciones, como puede observarse en la Figura 3. El trabajo de los científicos e ingenieros es calcular la función que, considerando un factor de error estimado, se ajuste mejor a las mediciones; por lo tanto, podría haber más de una curva de aproximación. Como se verá más adelante, el tipo y la cantidad de *funciones de aproximación* será proporcional a la cantidad y complejidad de los calculos realizados para encontrarlas.

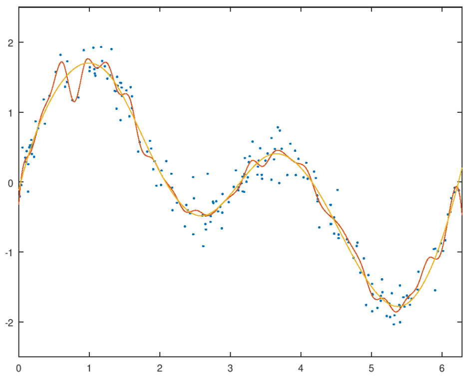


Figura 3

Regresión lineal por mínimos cuadrados

El método de regresión lineal por mínimos cuadrados es un método de interpolación que permite ajustar una función, a partir de un conjunto discreto de datos. El término *regresión* fue utilizado por primera vez por el estadista inglés Francis Galton para describir un fenómeno biológico en el que las estaturas de ancestros altos tendían a regresar hacia un promedio determinado. Por otro lado, como ya se mencionó, el término *lineal* se refiere a que la función que se ajustará al conjunto de datos será una recta. Mientras que *mínimos cuadrados* está relacionado con el cálculo del *mínimo* valor de la varianza; es decir, la suma del *cuadrado* de de los errores.

Ventajas del método de mínimos cuadrados

Este método permite tratar los residuos[1] como si fueran cantidades continuas, donde las derivadas[2] existan. Esto es de suma importancia para la posibilidad de predecir dónde podrían existir otros puntos en la función.

[1]En estadística, diferencias entre los valores de la variable dependiente observados y los valores que predecimos a partir de nuestra recta de regresión.

[2]Medida de cuánto cambia la salida de una función con respecto a su entrada.

Desventajas del método de mínimos cuadrados

El uso de valores alejados del conjunto de mediciones tiene un efecto desproporcional en el cálculo del método, si se busca encontrar la ecuación de una curva. Este se debe a que se utilizan los residuos en vez del valor absoluto de los residuos. Los valores alejados del conjunto de mediciones tienen constantemente grandes residuos y afectarán más a la función calculada que los puntos incluso más cercanos a ésta.

Tipos de regresiones por mínimos cuadrados

En estadístistica, el más común de los tipos de ajustes por mínimos cuadrados el cálculo de regresión lineal, a partir de un conjunto de observaciones.

Los mínimos cuadrados también son utilizados para ajustar a parámetros no lineales. Sin embargo, esta técnica se se torna de una gran complejidad.

Cálculo de la regresión lineal por mínimos cuadrados

Dado un conjunto de observaciones $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, se busca un recta que cumpla

$$y = a_0 + a_1 x + e$$

donde a_0 y a_1 son respectivamente el coeficiente de la intersección con el eje y y la pendiente de la recta; asimismo, e es el error o diferencia entre la función ajustada y las observaciones. A partir de esta ecuación, es que puede calcularse un error de la siguiente manera:

$$e = y - a_0 - a_1 x$$

Para ajustar la recta a través de los datos con que se cuenta, el método minimiza la suma de los cuadrados de los residuos entre la y medida y la y calculada; es decir, a partir de la siguiente ecuación, donde S_r es la suma de los residuos, y n es la cantidad de mediciones u observaciones con que se cuentan

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_{i\ medida} - y_{i\ calculada})^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

se calculan las derivadas parciales con respecto a a_0 y a a_1, y se igualan a cero:

$$\frac{\partial S_r}{\partial a_0} = \frac{\partial}{\partial a_0} \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 = -2 \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) = 0$$

$$\frac{\partial S_r}{\partial a_1} = \frac{\partial}{\partial a_1} \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 - 2 \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) x_i = 0$$

De lo anterior, se obtienen las siguientes ecuaciones:

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a_0 - \sum_{i=1}^{n} a_1 x_i = 0$$

$$\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} a_0 x_i - \sum_{i=1}^{n} a_1 x_i^2 = 0$$

Dado que $\sum_{i=1}^{n} a_0 = na_0$, resulta el siguiente sistema de ecuaciones:

$$na_0 + \sum_{i=1}^{n} a_1 x_i = \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} a_0 x_i + \sum_{i=1}^{n} a_1 x_i^2$$

Resolviéndose el sistema anterior, se tiene que:

$$a_0 = \frac{\sum_{i=1}^{n} y_i}{n} - a1 \frac{\sum_{i=1}^{n} x_i}{n}$$

es decir,

$$\boldsymbol{a_0} = \overline{y} - a_1 \overline{x}$$

y

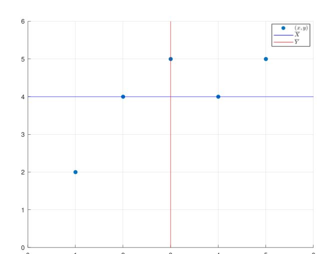
$$\boldsymbol{a_1} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Donde $\overline{y}$ y $\overline{x}$ son respectivamente los valores de la media de y y de x. Así pues, con las fórmulas anteriores puede calcularse la recta $\boldsymbol{y = a_0 + a_1 x}$, que es la aproximación que se buscaba.

Ejemplo numérico

Considerando los datos de la siguiente tabla como los valores de las mediciones u observaciones, pueden graficarse de la siguiente manera:

x	y
1	2
2	4
3	5
4	4
5	5



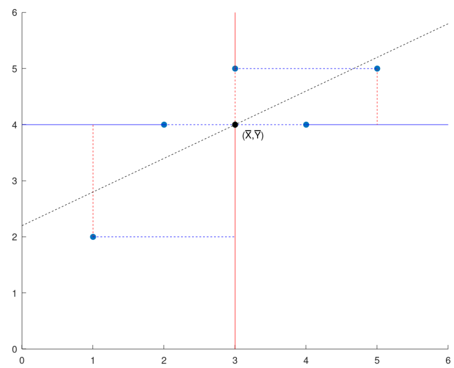
Donde

$$\boldsymbol{\overline{x}} = \frac{\sum x_i}{n} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = \mathbf{3}$$

$$\boldsymbol{\overline{y}} = \frac{\sum y_i}{n} = \frac{2+4+5+4+5}{5} = \frac{20}{5} = \mathbf{4}$$

son la media de los valores x_i ($\overline{x}$) y la media de los valores y_i ($\overline{y}$).

Note que se trazaron dos líneas, cuya intersección es $(\overline{x}, \overline{y}) = (3, 4)$, porque la recta de regresión lineal debe pasar por el punto $(\overline{x}, \overline{y}) = (3, 4)$, y que para ajustar una función a los datos de la tabla, es necesario calcular sus distancias hacia estas coordenadas, como semuestra en la siguiente gráfica:



Es decir, a partir de los datos y de los valores medios calculados, tenemos:

x_i	$\overline{x}$	$x_i - \overline{x}$	y_i	$\overline{y}$	$y_i - \overline{y}$
1	3	-2	2	4	-2
2	3	-1	4	4	0
3	3	0	5	4	1
4	3	1	4	4	0
5	3	2	5	4	1

Como el método utiliza el cuadrado de las diferencias de $x_i - \overline{x}$, y el producto de las diferencias de $(x_i - \overline{x})(y_i - \overline{y})$ para calcular la pendiente (a_1) de la recta $y = a_0 + a_1 x$, se tiene:

$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$
-2	4	-2	4
-1	1	0	0
0	0	1	0
1	1	0	0
2	4	1	2
$\sum (x_i - \overline{x})^2 =$	10	$\sum (x_i - \overline{x})(y_i - \overline{y}) =$	6

Por lo tanto, dado que

$$\boldsymbol{a_1} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{6}{10} = \mathbf{0.6}$$

recordando que $\overline{x} = 3$ y $\overline{y} = 4$, y

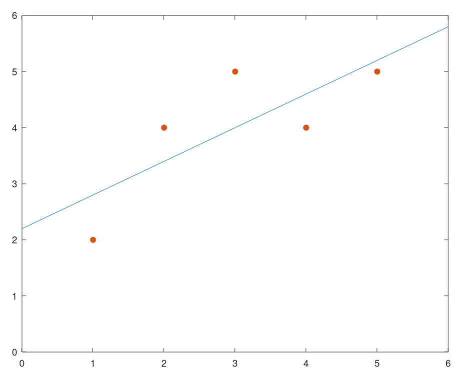
$$\boldsymbol{a_0} = \overline{y} - a_1\overline{x} = 4 - (0.6)3 = \mathbf{2.2}$$

entonces, por el método de regresión lineal por mínimos cuadrados, se ajusta el conjunto de datos a la recta

$$y = a_0 + a_1 x$$

$$\boldsymbol{y = 2.2 + 0.6x}$$

que se muestra en la gráfica siguiente



Gracias a esta ecuación, puede interpolarse (ajustarse, predecirse) posibles valores para y, a partir de valores de x que no hayan sido medidos u observados.