Regresión polinomial de segundo orden por el método de mínimos cuadrados

Alfredo Villagrán Olguín

Mayo, 2020

Regresión polinomial

Como se explicó en la nota sobre interpolación, es posible ajustar una serie de observaciones o mediciones a una curva. Cuando el ajuste se da a una línea recta, se le llama regresión lineal; cuando la curva no es lineal, y es posible, la regresión es llamada polinomial. Esto último, debido a que la curva pueda representarse a través de un polinomio

$$y = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n + e$$

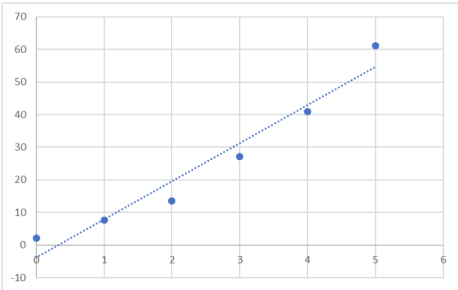


Figura 1

La Figura 1 muestra una serie de datos que son ajustados a una línea recta, a través de regresión lineal; mientras que la Figura 2 muestra los mismos datos ajustados a una función cuadrática, a través de una regresión polinomial de segundo grado.

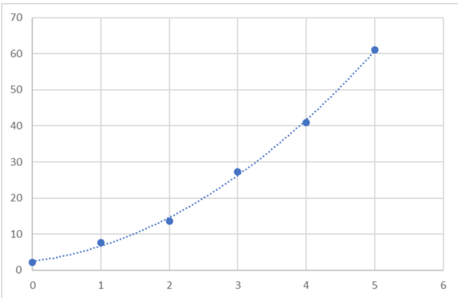


Figura 2

Es necesario recordar que el método de aproximación por mínimos cuadrados se depende de las siguientes bases estadísticas:

1. Cada x tiene una magnitud fija que no es aleatoria y se conoce desde un principo del procedimiento.
2. Los valores de y son independientes, aleatorios y con la misma varianza.
3. Los valores de y para cada x necesitan estar distribuidos normalmente.

Cálculo de regresión polinomial de segundo orden

Es posible extender el método de regresión lineal por mínimos cuadrados para ajustarlo a un polinomio de grado superior. A continuación se muestra como ajustar una serie de n datos a una función cuadrática, es decir, a un polinomio de segundo grado.

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

$$y = a_0 + a_1 x + a_2 x^2 + e$$

A partir de esta ecuación, es que puede calcularse un error de la siguiente manera:

$$e = y - a_0 - a_1 x - a_2 x^2$$

Para ajustar el polinomio 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

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_{i\ medida} - y_{i\ calculada})^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$$

se calculan las derivadas parciales con respecto a a_0 a_1, y a a_2, y se igualan a cero:

$$\frac{\partial S_r}{\partial a_0} = \frac{\partial}{\partial a_0} \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2 = -2 \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0$$

$$\frac{\partial S_r}{\partial a_1} = \frac{\partial}{\partial a_1} \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2 = -2 \sum_{i=1}^{n} x_i (y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0$$

$$\frac{\partial S_r}{\partial a_2} = \frac{\partial}{\partial a_1} \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2 = -2 \sum_{i=1}^{n} x_i^2 (y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0$$

De lo anterior, se obtienen las siguientes ecuaciones lineales con tres incógnitas, donde $na_0 = \sum_{i=1}^{n} a_0$:

$$na_0 + (\sum_{i=1}^{n} x_i)a_1 + (\sum_{i=1}^{n} x_i^2)a_2 = \sum_{i=1}^{n} y_i$$

$$(\sum_{i=1}^{n} x_i)a_0 + (\sum_{i=1}^{n} x_i^2)a_1 + (\sum_{i=1}^{n} x_i^3)a_2 = \sum_{i=1}^{n} x_i y_i$$

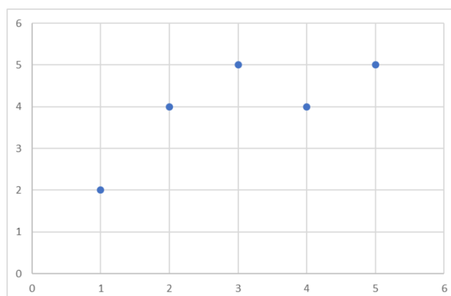
$$(\sum_{i=1}^{n} x_i^2)a_0 + (\sum_{i=1}^{n} x_i^3)a_1 + (\sum_{i=1}^{n} x_i^4)a_2 = \sum_{i=1}^{n} x_i^2 y_i$$

Así pues, con las fórmulas anteriores y a través de alguno de los métodos de resolución de sistemas de ecuaciones vistos al principio del curso, puede calcularse la curva $\boldsymbol{y = a_0 + a_1 x + a_2 x^2}$, que es la aproximación que se buscaba.

Ejemplo numérico

Considerando los datos de la siguiente tabla como los valores de las mediciones u observaciones, cuya grafica se muestra a un costado

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



pueden calcularse:

$$\sum \boldsymbol{y_i} = 2 + 4 + 5 + 4 + 5 = \mathbf{20}$$

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

$$\sum \boldsymbol{x_i^2} = 1 + 4 + 9 + 16 + 25 = \mathbf{55}$$

$$\sum \boldsymbol{x_i^3} = 1 + 8 + 27 + 64 + 125 = \mathbf{225}$$

$$\sum \boldsymbol{x_i^4} = 1 + 16 + 81 + 256 + 625 = \mathbf{979}$$

$$\sum \boldsymbol{x_i y_i} = 2 + 8 + 15 + 16 + 25 = \mathbf{66}$$

$$\sum \boldsymbol{x_i^2 y_i} = 2 + 16 + 45 + 64 + 125 = \mathbf{252}$$

Por lo tanto, de

$$na_0 + (\sum_{i=1}^{n} x_i)a_1 + (\sum_{i=1}^{n} x_i^2)a_2 = \sum_{i=1}^{n} y_i$$
$$(\sum_{i=1}^{n} x_i)a_0 + (\sum_{i=1}^{n} x_i^2)a_1 + (\sum_{i=1}^{n} x_i^3)a_2 = \sum_{i=1}^{n} x_i y_i$$
$$(\sum_{i=1}^{n} x_i^2)a_0 + (\sum_{i=1}^{n} x_i^3)a_1 + (\sum_{i=1}^{n} x_i^4)a_2 = \sum_{i=1}^{n} x_i^2 y_i$$

se tiene el siguiente sistema lineal de 3 ecuaciones con 3 incógnitas:

$$\begin{bmatrix} 5 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 20 \\ 66 \\ 252 \end{bmatrix}$$

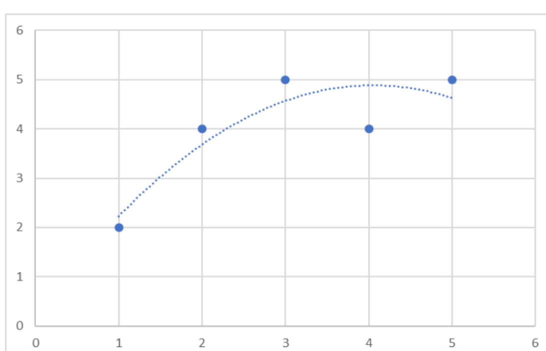
cuya solución, a través de alguno de los métodos estudiados al inicio del semestre, es:

$$\boldsymbol{a_0} = 0.2$$
$$\boldsymbol{a_1} = 2.314286$$
$$\boldsymbol{a_2} = -0.285714$$

Entonces, por el método de regresión lineal por mínimos cuadrados, se ajusta el conjunto de datos a la curva

$$y = a_0 + a_1 x + a_2 x^2$$
$$\boldsymbol{y = 0.2 + 2.314286x - 0.285714x^2}$$

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.