## Regresión lineal. Método de ajuste por mínimos cuadrados a una recta.

La chicharra *(Quesada gigas)* es un insecto hemíptero que tiene un abdomen cónico en cuya base los machos disponen de un sistema con el cual producen un ruido estridente y monótono. La frecuencia de su canto establece una relación lineal con la temperatura en el sentido de que, cuanto mayor es la temperatura, tanto mayor es la frecuencia de su canto. Por el contrario, cuando la temperatura disminuye también lo hace el número de chirridos. Lo interesante de esta conclusión es que si contamos el número de chirridos por unidad de tiempo, dispondremos de un termómetro que nos aporta información sobre la temperatura ambiente.

Imaginemos que vamos al campo y hacemos las medidas de frecuencia y temperatura mencionadas. Los datos obtenidos[i] podrían ser parecidos a los que se muestran en la Figura 1.

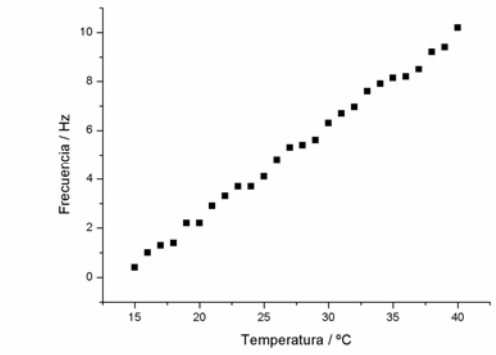


**Figura 1**

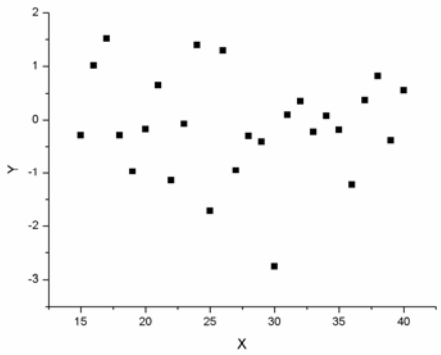


**Figura 2**

Efectivamente, se observa que la representación de la frecuencia frente a la temperatura se distribuye más o menos a lo largo de una línea recta. Dicho de otra manera, se observa un comportamiento más o menos lineal. No obstante, "más o menos lineal" no es una expresión muy objetiva, todo lo contrario. Para transformar un concepto ambiguo en algo objetivo, mensurable y comparable numéricamente, es para lo que recurrimos al concepto de regresión lineal.

El método de regresión lineal, también conocido como el método de ajuste por mínimos cuadrados a una recta, sirve para encontrar la recta que mejor se aproxima a los puntos experimentales. Matemáticamente una línea recta queda definida por la expresión,

$$Y = A + B \cdot X$$

Donde Y representa los valores mostrados en el eje vertical (ordenadas) mientras que la variable X corresponde con el eje horizontal (abscisas). Los parámetros *A* y *B*, son los que definen inequívocamente a la recta, es decir, dos rectas son diferentes o dicho de otro modo, dos relaciones lineales son distintas si difieren en el valor de alguno de estos dos parámetros.

*A* es el término independiente, denominado *ordenada en el origen* porque corresponde con el valor de Y cuando X vale cero y *B* es la pendiente. Informa sobre la inclinación de la recta y corresponde con la tangente del ángulo que forma la recta con la horizontal. Cuando se trata de una recta horizontal *B* vale cero ($B = \tan(0º) = 0$), e indica que la variable Y es independiente de X. Cuando la recta es vertical, el parámetro *B* no está definido ($B = \tan(90º)$).

¿En qué consiste el método de regresión lineal? En calcular la distancia más corta entre cada punto experimental y la recta hipotética que se busca. La recta más representativa es la que hace mínima estas distancias. El procedimiento es sencillo pero laborioso, no obstante, los cálculos están implementados en casi todas las calculadoras científicas y en aplicaciones informáticas básicas como las hojas de cálculo (Excel, OpenOffice.org Calc, etc.). Bastará con proporcionar los datos de X e Y y la aplicación hará el resto. De forma casi instantánea nos proporcionará los parámetros *A* y *B* que identifica la recta que buscamos.

Hay algo importante a tener en cuenta a la hora de aplicar el método. ¡Siempre se puede aplicar! A cualquier relación de datos X e Y. Tanto a relaciones como la mostrada en la Figura 1, como a conjuntos de datos como los que se observan en la Figura 2, que no tienen pinta de ajustarse linealmente. Y una vez aplicado, el método SIEMPRE nos dará como resultado los parámetros que definen la recta que mejor se aproxima a los puntos.

Pero esto no significa que los datos se ajusten linealmente. Para saber cuando el ajuste es bueno o no, además de los mencionados parámetros *A* y *B*, el método nos proporciona también el coeficiente de regresión, *r*.

El valor de *r* está comprendido entre –1 y +1. El valor +1, indica que existe una relación lineal de tal manera que cuando aumenta la variable X también lo hace la Y. Si se obtiene un coeficiente *r* = -1, la relación entre los datos también es lineal; el signo indica aquí que la relación entre X e Y es inversamente proporcional, es decir, que mientras la X aumenta, Y disminuye.

Cuanto más próximo sea el coeficiente *r* a cualquiera de estos dos valores (+1 ó -1), más ajustada es la distribución de los datos a una relación lineal. Pero si *r* vale cero o un valor próximo, entonces debemos descartar la existencia de una relación lineal.

A veces, como en el ejemplo que se muestra a continuación, la aplicación no proporciona el valor de *r* sino el de $r^2$, en cuyo caso la discusión sobre su valor se ciñe al intervalo 0 – 1.

Apliquemos el método de regresión lineal a los datos mostrados en las Figuras 1 y 2.

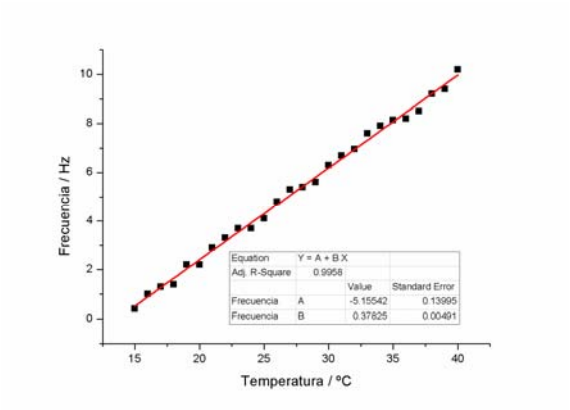


**Figura 3**

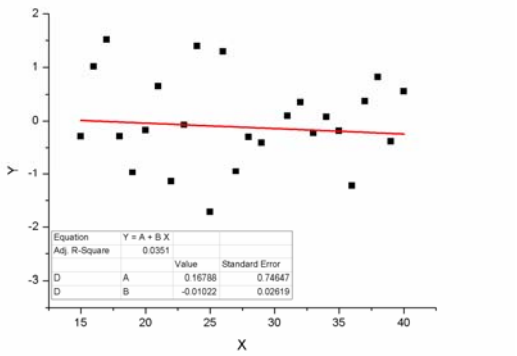


**Figura 4**

En la Figura 3 se observa la recta que mejor se aproxima a la relación que, ya a principios de este escrito, anunciábamos lineal (Figura 1). La confirmación de este hecho lo da el valor de $r^2$ (Adj. R-Square en la figura) que es tan cercano a 1 como 0.9958.

La Figura 4 muestra el resultado obtenido al aplicar el método de regresión lineal sobre la relación de puntos mostrada en la Figura 2, una relación que a simple vista, no parece muy lineal. El método, no obstante, aporta los datos de la recta encontrada así como el del coeficiente de regresión. Un valor tan pobre como $r^2 = 0.0351$.

En atención a los valores de $r^2$ obtenidos en ambos ajustes concluimos que en el caso de la Figura 3, la recta obtenida es representativa de los datos y puede utilizarse en publicaciones, informes y cálculos como una expresión equivalente a los datos experimentales. Pero en el caso de la Figura 4, con un $r^2$ de 0.0351, lo único que podemos afirmar es que los puntos no se distribuyen linealmente, por lo tanto, la recta obtenida por el método, no es representativa de los datos y debe arrojarse al más oscuro de los olvidos.

[i] Los datos son inventados.