Regresión lineal por cuadrados mínimos

Regresión lineal por cuadrados mínimos ordinarios

Supongamos que, a través de una serie de mediciones, se han determinado un conjunto de n pares de valores de dos magnitudes físicas, X e Y, asociadas a un cierto fenómeno. Es decir, se tiene el siguiente conjunto de datos experimentales: $(x_i \pm \Delta x_i , y_i \pm \Delta y_i)$, con i de 1 hasta n. Supongamos además que hay motivos suficientes como para conjeturar que existe una relación funcional lineal entre X e Y (Figura 1, panel izquierdo), tal que:

$Y = \mathbf{a} + \mathbf{b}X$

donde **a** corresponde a la ordenada al origen y **b** a la pendiente de la recta.

En este caso, cabe preguntarse: ¿Me permiten mis datos afirmar que la relación lineal es correcta, al menos dentro del rango de valores medidos? En caso afirmativo, ¿cómo determino los parámetros **a** y **b**, y cuáles son sus incertezas?

La técnica más usual para responder esas preguntas se conoce como el método de regresión lineal por cuadrados mínimos ordinarios, el cual supone despreciables las incertezas Δx_i y Δy_i.

¿Cómo determinar a y b? Para cada par de datos medido u observado (x_i,y_i), se define la cantidad: $d_i = y_i - (a+bx_i)$, donde d_i es la distancia entre el valor experimental observado y_i y el valor predicho por la recta y_i^* para el valor experimental xi, es decir $y_i^*=a+bx_i$ (Figura 1, panel derecho).

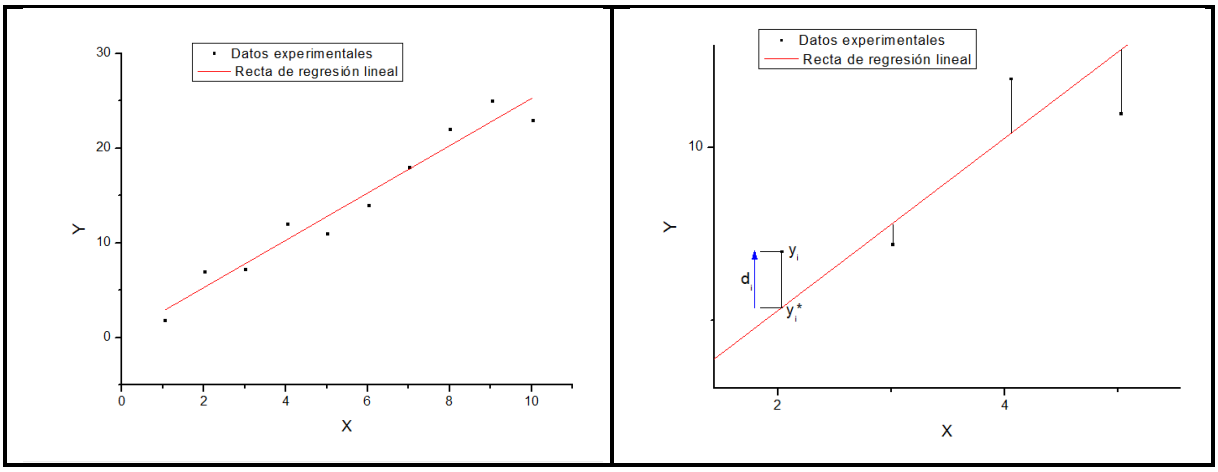

Figura 1: Panel izquierdo: datos experimentales y recta de regresión lineal. Panel derecho: visualización de $d_i = y_i – y_i^*$.

Es intuitivo ver que, si la relación es efectivamente lineal y los **a** y **b** son los correctos, cada d_i debe ser "pequeño". Como criterio para satisfacer esto, se puede proponer que la suma de todos los d_i sea mínima. Sin embargo, la suma de los d_i no es buen criterio, porque los d_i pueden ser positivos y negativos y cancelarse mutuamente. Para evitar este problema, se pueden sumar los módulos o, lo que es más conveniente, los d_i elevados al cuadrado: d_i^2. En el método de regresión lineal por **cuadrados mínimos ordinarios** se eligen los valores de **a** y **b** tales que minimicen la suma de los cuadrados de los d_i:

$$\Sigma' = \sum_{i=1}^{n} d_i^{\,2}$$

Es decir, la técnica de cuadrados mínimos ordinarios nos da expresiones analíticas para calcular la ordenada al origen **a** y la pendiente **b** de la recta de regresión lineal y una expresión para sus incertezas Δa y Δb respectivamente.

Queda todavía por contestar la pregunta de si los datos permiten asumir que las variables X e Y guardan una relación lineal. La técnica de regresión lineal por cuadrados mínimos ordinarios nos da un criterio para evaluar la confiabiliad o calidad de la relación lineal y es a través de un coeficiente, **r**, llamado coeficiente de correlación lineal de los datos.

El valor del índice de correlación lineal **r** varía en el intervalo [-1,1], indicando el signo el sentido de la relación:

- Si **r = 1**: **correlación lineal perfecta positiva** y los valores predichos coinciden con los observados, ya que todos los puntos de la nube están en la recta. Es decir, existe dependencia funcional que viene reflejada por una recta creciente.
- Si **r = -1**, la **correlación lineal es perfecta negativa** y, aquí también, los valores predichos coinciden con los observados, pero la recta es decreciente. De nuevo es un caso de dependencia funcional.
- Si **r = 0**, la **correlación lineal es nula**. Es decir, no hay asociación lineal y por mucho que varíe X, la variable Y no se verá afectada (de forma lineal).
- Si **-1 < r < 0**, la **correlación lineal será negativa** y la recta será decreciente puesto que el signo r coincide con el de la pendiente. Si r es cercano a 0 diremos que la relación es débil, y cuanto más se acerque a -1 consideráremos que la relación es más fuerte.
- Si **0 < r < 1**, la **correlación lineal es positiva**. Esto indica que la recta es creciente y cuando los valores de una variable crecen lo de la otra también crecerán. Consideraremos también que cuanto más se acerque a 0 más débil es la relación entre las variables y si el valor es próximo a 1 la relación podrá considerarse fuerte.

Regresión lineal por Cuadrados mínimos ponderados.

Se tiene un conjunto de datos experimentales: ($x_i \pm \Delta x_i$, $y_i \pm \Delta y_i$), con i de 1 hasta n, y tal como en el caso anterior supongamos además que hay motivos suficientes como para conjeturar que existe una relación funcional lineal entre X e Y, tal que:

$Y = \mathbf{a} + \mathbf{b}X$

donde **a** corresponde a la ordenada al origen y **b** a la pendiente de la recta.

Antes de ver como se obtienen **a** y **b** usando un proceso de minimización equivalente al del caso anterior, vamos a dar un paso más para decidir que suma minimizar. Para ello, vamos a suponer que los errores de la magnitud X son despreciables frente a los errores de Y. Nótese, de paso, que d_i (Figura 2, panel derecho) es la distancia, en la dirección del eje y, entre y_i y el valor predicho por la recta $y_i^* = \mathbf{a} + \mathbf{b}x_i$. Es claro que, si los x_i "no tienen incerteza", la relación es lineal y sus parámetros son **a** y **b**, entonces las diferencias entre y_i e y_i^* son atribuibles en parte a las incertezas Δy_i. Pero si estas incertezas son distintas para cada y_i (Figura 2, panel izquierdo), parece sensato realizar el procedimiento de minimización asignando una mayor importancia a los d_i provenientes de valores de los y_i que tengan errores más chicos.

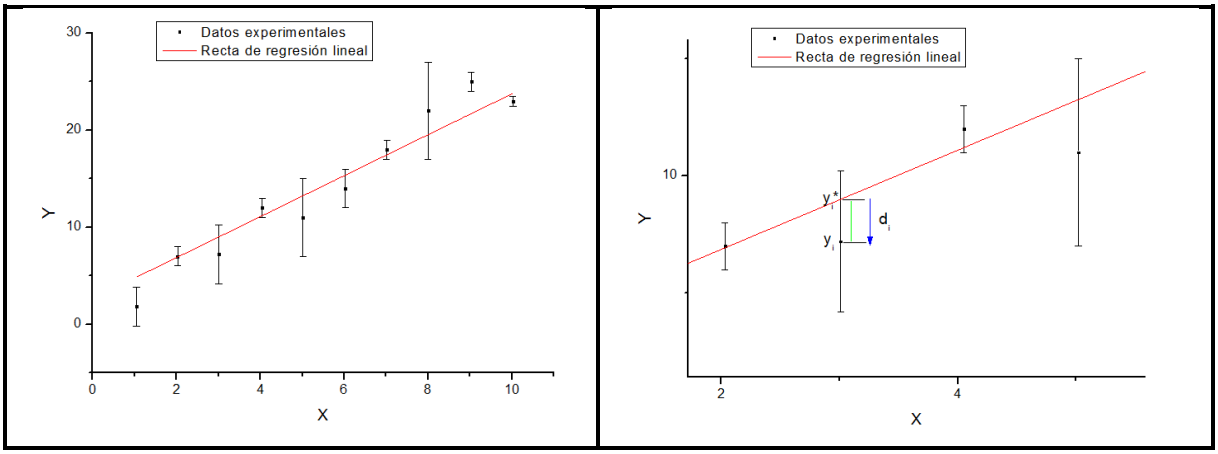

Figura 2: Panel izquierdo: datos experimentales con incerteza en Y, Δy_i, y recta de regresión lineal. Panel derecho: zoom en el que se visualiza d_i.

Esto se logra por el procedimiento de "**ponderar**", que consiste en multiplicar a cada d_i por una cantidad que sea mayor cuanto más pequeño sea el correspondiente Δy_i. La forma más sencilla de hacer esto es dividir a cada d_i por Δy_i, tal que cuanto menor es la incerteza de y_i, mayor es el coeficiente que tiene el d_i. Usando este criterio, lo que se debe minimizar es:

$$\Sigma = \sum_{i=1}^{n} \left(\frac{d_i}{\Delta y_i} \right)^2 = \sum_{i=1}^{n} \left(\frac{y_i - a - bx_i}{\Delta y_i} \right)^2$$

Al método de obtención de **a** y **b** usando la minimización de Σ se lo llama "**cuadrados mínimos ponderados**".

Al igual que la técnica de cuadrados mínimos ordinarios, la técnica de cuadrados mínimos ponderados nos da expresiones analíticas para calcular la pendiente y la ordenada de la recta de regresión lineal y expresiones analíticas para sus incertezas: **a** +/- Δa y **b** +/- Δb. La técnica de regresión lineal por cuadrados mínimos ponderados nos da también un criterio para evaluar la confiabilidad o calidad de la relación lineal y es a través del mismo coeficiente, **r**, llamado coeficiente de correlación lineal de los datos.

Existen Paquetes estadísiticos y/o programas de análisis de datos tales como el Origin, que hacen estas cuentas por nosotros y nos arrojan los valores de **a** +/- Δa, **b** +/- Δb y **r**. (Consultar el tutorial de Origin o SciDAVis sobre gráficos de dispersión y regresión lineal).

Por último: al realizar una regresión lineal por cuadrados mínimos, recuerde colocar siempre como variable **Y** aquella que posee el mayor error relativo.