AJUSTE POR CUADRADOS MÍNIMOS

1. Método de los cuadrados mínimos

Hasta ahora, hemos estudiado como expresar correctamente los errores medidos de forma directa e indirecta, analizando las fuentes de errores, procurando entender cómo llevar correctamente a cabo cualquier proceso de medición. Al calcular una magnitud de forma indirecta, la fórmula que relacionaba las variables directas con la indirecta era conocida previamente.

Cuando se estudia un fenómeno cualquiera, no se conoce a priori la relación entre las variables. La física real comienza cuando se estudia la interdependencia entre dos o más magnitudes con el fin de establecer las leyes físicas que permitan describir la evolución del sistema en estudio. Para esto, en Física, se cuenta con la inestimable ayuda de las matemáticas.

Desafortunadamente, en la práctica, las magnitudes que se miden están afectadas por los errores de medición (que se recuerda no pueden eliminarse), por lo que se hace necesaria alguna estrategia que permita obtener la relación más probable entre las magnitudes físicas vinculadas. El caso más sencillo es que las cantidades estén relacionadas de forma lineal y eso es lo que se tratará en este apunte.

Como primera medida, se supone que la serie de valores (x, y) de las variables que se miden experimentalmente para verificar una ley física, son en principio independientes entre sí. Como consecuencia, sus incertezas son también independientes. Así, se tienen N pares de valores (x_i, y_i). ¿Cuál es la variable independiente y cuál la dependiente? Para definir esto, se deben observar los errores de medición de cada uno. La que tenga menor error absoluto, será la variable independiente, y su error se considerará despreciable frente al error de la otra variable. El objetivo del método de ajuste por *cuadrados mínimos* será obtener una curva tal que la distancia vertical de los valores experimentales a la misma sea mínima.

Se parte de la suposición de que los valores experimentales se relacionan entre sí de forma lineal:

$$y = ax + b \qquad (1.1)$$

El problema consiste en determinar el valor de los parámetros a y b. Si todos los puntos estuvieran en una única recta, entonces se cumpliría:

$$y_i - ax_i + b = 0 \qquad \forall\, i$$

Pero como cualquier medición experimental contiene incerteza, entre ambos miembros de la ecuación 1.1 existe un residuo no nulo, al que llamaremos e_i:

$$y_i - ax_i + b = e_i \neq 0$$

e_i describe el error en el ajuste en el punto i-ésimo.

Si se obtienen N lecturas x_i, y_i, habrá N desviaciones o errores e_i. El valor de e_i dependerá en cada punto de los valores de los parámetros a y b. ¿Cómo se determinan dichos valores? El objetivo es minimizar todos los errores e_i. Si se minimizara cada e_i por separado, se obtendrían N valores de a y b, pero lo que se desea es obtener un único valor para cada parámetro a y b. En vez de minimizar los errores por separado, lo que debe minimizarse es la suma de los cuadrados de los errores (S):

$$S = \sum_{i=1}^{N} e_i^2$$

Observación: podría pensarse por qué no minimizar la suma de errores solamente (no los cuadrados). Los valores de e_i pueden ser positivos y/o negativos, por lo que podrían compensarse y no representar el error total del ajuste.

Para minimizar S, debe cumplirse la condición doble:

$$\left.\frac{\partial S}{\partial a}\right|_{a=\bar{a}} = 0 \quad ; \quad \left.\frac{\partial S}{\partial b}\right|_{b=\bar{b}} = 0$$

En el caso de una recta:

$$\left.\frac{\partial S}{\partial a}\right|_{a=\bar{a}} = \left.\frac{\partial}{\partial a}\sum(y_i - ax_i + b)^2\right|_{a=\bar{a}} = 2\sum(y_i - \bar{a}x_i + \bar{b})\,(-x_i) = 0$$

$$\left.\frac{\partial S}{\partial b}\right|_{b=\bar{b}} = \left.\frac{\partial}{\partial b}\sum(y_i - ax_i + b)^2\right|_{b=\bar{b}} = 2\sum(y_i - \bar{a}x_i + \bar{b})\,(-x_i) = 0$$

De donde surge el sistema de ecuaciones:

$$\bar{a}\sum x_i^2 + \bar{b}\sum x_i - \sum x_i y_i = 0$$

$$\bar{a}\sum x_i + \bar{b}N - \sum y_i = 0$$

Cuya solución para los valores de solución de $\bar{a}$ y $\bar{b}$ son:

$$\bar{a} = \frac{\sum x_i.\sum y_i - N\sum x_i y_i}{(\sum x_i)^2 - N\sum x_i^2} \quad ; \quad \bar{b} = \frac{\sum x_i.\sum x_i y_i - \sum x_i^2 \sum y_i}{(\sum x_i)^2 - N\sum x_i^2} \qquad (1.2)$$

Observación: A menudo, en las aplicaciones científicas o de ingeniería, existe una sola variable *dependiente* o *respuesta* (y) que no se controla en el experimento. Dicha respuesta depende del valor de una o más variables independientes (x o x_i), respectivamente. Estas se miden con un error insignificante, y a menudo realmente se controlan, es decir, se conocen anticipadamente ayudando a predecir el valor de y_i. Las variables independientes no son variables aleatorias

2. Errores de los parámetros

Los parámetros a y b son calculados mediante las ecuaciones 1.2, pero tienen intervalos de incerteza propios. Esto permite determinar cuan confiables son sus valores. Pero, ¿cómo se calculan sus incertezas? Partimos de la hipótesis de que las incertezas de x son despreciables frente a las de y. Supongamos ahora que se repiten M veces las mediciones de cada una de las y_i, por lo que los valores de y_i fluctuaran mientras que los de x_i no, obteniéndose un conjunto de valores:

$$y_{ik}(k = 1,2,3,\ldots\ldots,M) \qquad \text{para cada } x_i$$

Se puede considerar a $\bar{a}$ y $\bar{b}$ como cantidades medidas indirectamente, siendo y_i (con i = 1,2,3,….N) las N cantidades medidas directamente, cada una con su error cuadrático medio $\sigma(y_i)$. Entonces, se puede aplicar tanto a $\bar{a}$ como a $\bar{b}$ la fórmula para calcular el error medio cuadrático de N lecturas indirectas:

$$\sigma_a = \sqrt{\sum_{i=1}^{N}\left(\frac{\partial \bar{a}}{\partial y_i}\right)^2 \sigma^2(y_i)} \quad ; \quad \sigma_b = \sqrt{\sum_{i=1}^{N}\left(\frac{\partial \bar{b}}{\partial y_i}\right)^2 \sigma^2(y_i)}$$

Si las diferentes y_i han sido obtenidas en las mismas condiciones, todas tendrán prácticamente el mismo σ:

$$\sigma(y_1) = \sigma(y_2) = \ldots \sigma(y_i) = \ldots = \sigma(y)$$

Resultando entonces una simplificación:

$$\sigma_a = \sigma_y\sqrt{\sum_{i=1}^{N}\left(\frac{\partial \bar{a}}{\partial y_i}\right)^2} \quad ; \quad \sigma_b = \sigma_y\sqrt{\sum_{i=1}^{N}\left(\frac{\partial \bar{b}}{\partial y_i}\right)^2}$$

Aún resta calcular $(\partial\bar{a}/\partial y_i)$ y $(\partial\bar{b}/\partial y_i)$. Si se llama D al denominador común de las expresiones de $\bar{a}$ y $\bar{b}$, las derivadas resultan:

$$\frac{\partial \bar{a}}{\partial y_i} = \frac{1}{D}\left(\sum_{j=1}^{N} x_j - N x_i\right) \qquad \text{y} \qquad \frac{\partial \bar{b}}{\partial y_i} = \frac{1}{D}\left[\left(\sum_{j=1}^{N} x_j\right)x_i - \sum_{j=1}^{N} x_j^2\right]$$

Luego, la suma de los cuadrados para la primera de ellas es:

$$\sum_{i=1}^{N}\left(\frac{\partial \bar{a}}{\partial y_i}\right)^2 = \frac{1}{D^2}\sum_{i=1}^{N}\left[\left(\sum_{j=1}^{N} x_j - N x_i\right)^2\right] = \frac{1}{D^2}\sum_{i=1}^{N}\left[\left(\sum_{j=1}^{N} x_j\right)^2 - 2N x_i \sum_{j=1}^{N} x_j + N^2 {x_i}^2\right]$$

$$= \frac{1}{D^2}\left[N\left(\sum_{j=1}^{N} x_j\right)^2 - 2N\sum_{i=1}^{N} x_i \sum_{j=1}^{N} x_j + N^2\sum_{i=1}^{N} {x_i}^2\right]$$

Si se tiene presente que $\sum_{i=1}^{N} x_i = \sum_{j=1}^{N} x_j$ (pues solo difieren en el nombre del índice), entonces:

$$\sum_{i=1}^{N}\left(\frac{\partial \bar{a}}{\partial y_i}\right)^2 = \frac{1}{D^2}\left[-N\left(\sum_{j=1}^{N} x_j\right)^2 + N^2\sum_{i=1}^{N} {x_i}^2\right] = \frac{N(-D)}{D^2} = -\frac{N}{D}$$

Con un desarrollo similar se llega a que:

$$\sum\nolimits_{i=1}\left(\frac{\partial \bar{b}}{\partial y_i}\right)^2 = -\frac{\sum {x_i}^2}{D}$$

Por lo tanto, las expresiones para los errores quedan reducidas a:

$$\sigma_a = \sigma_y\sqrt{-\frac{N}{D}} \qquad \text{y} \qquad \sigma_b = \sigma_y\sqrt{-\frac{\sum {x_i}^2}{D}}$$

Normalmente no se tienen M lecturas para cada punto (x_i, y_i), para cada x_i sólo se lee y_i una vez o unas pocas veces de modo que no hay modo de calcular σ_y. Sin embargo, es posible estimar el valor sobre la base de la siguiente hipótesis: las desviaciones de las y_{jk} alrededor de su promedio y_i son del mismo orden que las diferencias e_i, entonces:

$$\sigma_y = \sqrt{-\frac{\sum {e_i}^2}{N}}$$

Con lo cual, las expresiones finales para los errores de los parámetros a y b resultan:

$$\sigma_a = \sqrt{-\frac{\sum {e_i}^2}{N}}\sqrt{-\frac{N}{D}} \cong \sqrt{\frac{\sum {e_i}^2}{-D}}$$

$$\sigma_b = \sqrt{-\frac{\sum {e_i}^2}{N}}\sqrt{-\frac{\sum {x_i}^2}{D}} \cong \sigma_a\sqrt{\frac{\sum {x_i}^2}{N}}$$

3. Diagramas de dispersión

La principal pregunta que debe responderse es: ¿cuándo es posible utilizar el método de cuadrados mínimos o regresión lineal? Si las cantidades observadas tienen una fuente común de variación, se dice que están correlacionadas. Para determinar si existe o no correlación entre dos variables, se construye un diagrama de puntos (*x,y*) en un sistema de ejes coordenados. Así se obtiene un conjunto de puntos esparcidos al azar denominado *diagrama de dispersión*.

Pueden presentarse varias situaciones:

a) Valores distribuidos simétricamente alrededor de los valores de $\bar{x}$ e $\bar{y}$, o sea, desde el punto de coordenadas $(\bar{x}, \bar{y})$, formando una "nube" aproximadamente circular. (Fig. 1a)

b) La distribución de puntos se aproxima a una curva con pequeñas fluctuaciones. (Fig. 1b)

c) Puntos formando una nube alrededor de una curva con fluctuaciones de cierta importancia. (Fig. 1c)

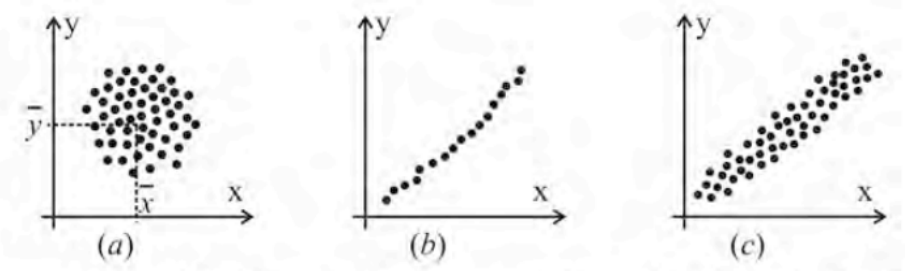

Figura 1. Distribución de puntos experimentales: a) circular, b) poca fluctuación, casi lineal, y c) con fluctuaciones de importancia.

En el caso a), las variables no están correlacionadas, no hay dependencia entre ellas. Los puntos están a distancias grandes de cualquier recta hipotética. No se cumple la hipótesis establecida.

En el caso b) existe dependencia casi total entre x e y. Existe correlación y es fuerte, hay una relación funcional $y = f(x)$. Las fluctuaciones se deben a incertezas casuales y de apreciación y pueden compensarse con un análisis estadístico.

En el caso c) existe una zona de dispersión amplia, pero las variables están vinculadas Los factores que las vinculan son variados y difíciles de aislar. No es interesante desde el punto de vista de la experimentación, ya que el diseño de la experiencia incorpora excesivas incertezas. Esto permite encontrar más de una recta, pero resulta imposible determinar cuál de todas es la correcta.

La curva que mejor ajusta los datos, se denomina *curva de regresión de y sobre x*. Si es una recta, se llama recta de regresión de y sobre x.

Claramente, el primer paso a realizar en el análisis de datos es el diagrama de dispersión, y luego de analizarlo, si existe correlación, se procede a realizar los cálculos de los parámetros o coeficientes de la recta a y b.

4. Coeficiente de correlación Pearson

Para dar una medida numérica que exprese cuan fuertemente depende una variable de otra se define el coeficiente de correlación de Pearson R o coeficiente de correlación muestral. El análisis de correlación intenta medir la fuerza de la relación lineal entre dos variables por medio de un sólo número. El coeficiente define por sí mismo el grado de asociación entre las variables seleccionadas, ya que la correlación entre variables es el grado de relación o conexión entre ellas.

Si "x_i" e "y_i" forman el conjunto de valores experimentales obtenidos, entonces, el coeficiente de correlación R está dado por la expresión:

$$R = \frac{S_{12}}{\sqrt{S_{11}S_{22}}}$$

Donde:

$$S_{12} = \sum x_i y_i - \left(\frac{(\sum x_i \sum y_i)}{N}\right)$$

$$S_{11} = \sum (x_i)^2 - \left[\frac{(\sum x_i)^2}{N}\right]$$

$$S_{12} = \sum (y_i)^2 - \left[\frac{(\sum y_i)^2}{N}\right]$$

donde "*N*" es el número de determinaciones efectuadas de "x" e "y".

El coeficiente de correlación entre dos variables se sitúa entre -1 y +1, inclusive. Si existe una relación lineal entre las variables el coeficiente es 1 o -1. El signo "-" indica que la relación lineal tiene pendiente negativa. El valor 0 (cero) refleja la ausencia de una relación lineal pero no de asociación entre las variables. En este caso se dice que las variables no están correlacionadas (es decir, son independientes). Toma valores intermedios si las variables están correlacionadas (pero no puede decirse nada del tipo de relación que las asocia); los valores serán más altos cuanto más fuerte sea la correlación. Se puede asegurar que los valores cercanos a +1 o -1 indican una tendencia lineal.

Es importante recordar que el coeficiente de correlación entre dos variables es una medida de su grado de linealidad.