MÉTODO DE AJUSTE POR REGRESIÓN LINEAL

Por lo general no es suficiente con saber medir una magnitud física y valorar el resultado de la medición desde su significado estadístico, la mayoría de las veces es necesario medir dos o más magnitudes y estudiar la interdependencia causal entre ellas.

Para establecer qué ley física rige en un sistema dado, primero hay que encontrar experimentalmente si existe algún tipo de relación entre los valores numéricos de las magnitudes intervinientes. Como primera aproximación se realiza una representación gráfica de los N pares de datos (X_i, Y_i) y se trata de descubrir cuál es la curva que sugiere la nube de puntos. El criterio de mínimos cuadrados reemplaza el juicio personal de quien mire los gráficos y define cuál es la mejor curva que se ajusta a los datos.

La función que relaciona dos variables X e Y puede ser lineal, exponencial, polinómica, etc. En este apunte vamos a centrarnos en una posible relación lineal entre dos variables: si la curva que mejor ajusta es una recta, se dice que existe "correlación lineal" entre las dos magnitudes medidas. En los programas como Excel, Origin, etc., este cálculo se realiza usando la herramienta "**regresión lineal**" o "estimación lineal".

Supongamos que se obtuvieron N puntos experimentales (Xi ,Yi) que al ubicarlos en la gráfica Y(X) parecen estar sobre una recta que se puede representar según la ecuación:

$$Y(X) = A \cdot X + B \qquad (1)$$

donde A y B son dos parámetros, a determinar, que caracterizan la pendiente de la recta y la ordenada al origen respectivamente. En el procedimiento denominado regresión lineal el objetivo es hallar el par de valores A y B, por método de cuadrados mínimos, que definen a la recta que mejor ajusta a los datos experimentales.

Si todos los puntos estuvieran exactamente sobre una recta se cumpliría que:

$$Y_i - (A \cdot X_i + B) = 0 \quad i = 1,2 \ldots N \qquad (2)$$

sin embargo, dado que los datos experimentales siempre están afectados de un error, la diferencia en la ecuación (2) no es cero, sino que para cada par de puntos (X_i,Y_i) se observa una diferencia $\boldsymbol{e_i}$

$$Y_i - (A \cdot X_i + B) = e_i \quad i = 1,2 \ldots N \qquad (3)$$

Dicho de otra forma, dado un par de valores A y B, de las N lecturas (X_i, Y_i) se obtendrán N desviaciones e_i. Si se asume que los errores están en su totalidad en los valores de Y_i, la desviación e_i se denomina "desviación vertical" y corresponde a la diferencia entre el valor medido Y_i y el valor de Y calculado $Y(X_i)$ (ver Fig. 1)

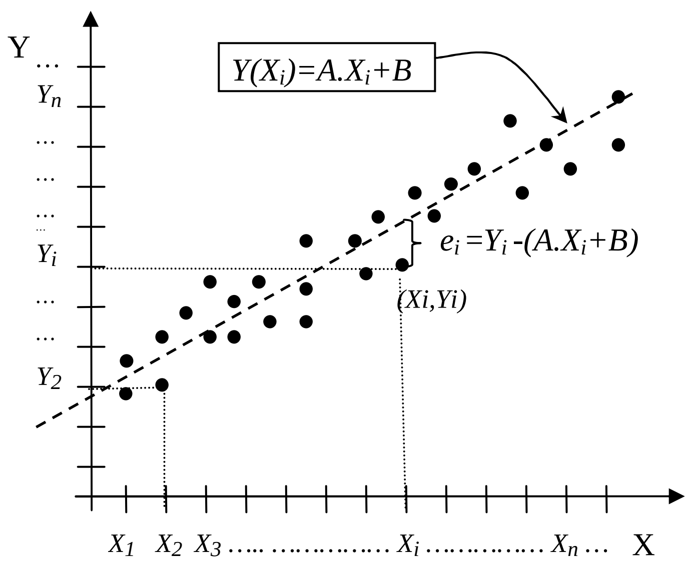


Figura 1. Representación gráfica de una serie de datos experimentales (Xi,Yi) y de una de las rectas $Y(X_i)=A.X_i+B$ que podría ajustar bien a los puntos.

Se considera como valor más acertado para asignar a los parámetros *A y B*, aquel que haga mínima la suma de los cuadrados de las desviaciones χ^2 ("Chi cuadrado"),

$$\chi^2 = \sum_{i=1}^{N} e_i^2 \Rightarrow Mínimo \qquad (4)$$

Observación: *esto es equivalente a pedir que sea máxima la probabilidad de que las desviaciones verticales, que efectivamente aparecen en la medición, sean mínimas.*

La condición de valor mínimo para la ecuación (4) se obtiene pidiendo que las derivadas parciales respecto de los parámetros *A y B* sean cero,

$$\frac{\partial \chi^2}{\partial A} = 0 \quad y \quad \frac{\partial \chi^2}{\partial B} = 0 \qquad (5)$$

En particular para este caso:

$$\frac{\partial \chi^2}{\partial A} = \sum_{i=1}^{N} \frac{\partial (Y_i - AX_i - B)^2}{\partial A} = 2 \cdot \sum_{i=1}^{N} \left[(Y_i - AX_i - B)(-X_i)\right] = 0$$

$$\frac{\partial \chi^2}{\partial B} = \sum_{i=1}^{N} \frac{\partial (Y_i - AX_i - B)^2}{\partial B} = 2 \cdot \sum_{i=1}^{N} \left[(Y_i - AX_i - B)(-1)\right] = 0$$

de donde resulta el sistema de ecuaciones:

$$A\sum_{i=1}^{N} (X_i)^2 + B\sum_{i=1}^{N} (X_i) - \sum_{i=1}^{N} (Y_i \cdot X_i) = 0 \qquad (6)$$

$$A\sum_{i=1}^{N} (X_i) + B \cdot N - \sum_{i=1}^{N} (Y_i) = 0 \qquad (7)$$

Las ecuaciones (6) y (7) pueden expresarse también

$$A \cdot \overline{X^2} + B \cdot \overline{X} - \overline{XY} = 0 \qquad (8)$$

$$A \cdot \overline{X} + B + \overline{Y} = 0 \qquad (9)$$

donde $\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$ $\quad \overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$ $\quad \overline{XY} = \frac{\sum_{i=1}^{N} X_i \cdot Y_i}{N}$ $\quad \overline{X^2} = \frac{\sum_{i=1}^{N} X_i^2}{N}$

La solución de este par de ecuaciones da los valores buscados de *A y B*, es decir, el de aquellos que cumplen con la condición de la ecuación (4).

La ecuación (9) nos dice que la mejor recta pasa por $\overline{X}$ e $\overline{Y}$ [1], es decir que el punto $G = \left(\overline{X}, \overline{Y}\right) = \left(X_c, Y_c\right)$ pertenece a la recta que buscamos.

Resolviendo las Ecuaciones (8) y (9), obtenemos los valores buscados para la ordenada al origen, *B*, y la pendiente *A* de la recta que mejor ajusta a nuestros datos experimentales:

$$A = \frac{\overline{XY} - \overline{X}\ \overline{Y}}{\overline{X^2} - \overline{X}^2} \quad (10) \qquad B = \frac{\overline{X^2}\ \overline{Y} - \overline{X}\ \overline{XY}}{\overline{X^2} - \overline{X}^2} \quad (11)$$

[1] El valor promedio de una variable $\overline{X}$ también podrá verlo expresado como $\left\langle X \right\rangle$

► *¿Qué intervalo de incerteza corresponde asignar a los valores de A y de B determinados?*

Como sabemos, cada medición tiene asociado un intervalo de incerteza σ. En consecuencia, cada par de datos *(X_i, Y_i)* también lo tiene y eso redunda en una incerteza para los valores de *A* y *B* calculados.

La mayoría de los programas de cálculo indican los valores de A y B, pero no de sus incertezas asociadas. Para conocer el intervalo de incertidumbre de cada par *(X_i, Y_i)* deberíamos medirlo $M \times N$ (*M* veces cada valor de X_i y *N* veces cada valor de Y_i). En la práctica eso no se hace y en compensación se realizan diversas aproximaciones. La primera es que la desviación estándar en los valores medidos de *X* es mucho menor que la desviación estándar en los valores medidos de *Y* $(\sigma_X << \sigma_Y)$, por lo cual se considera despreciable. Luego, aplicando la relación que permite calcular la desviación estándar de una magnitud que está dada en función de otras variables medidas se llega a la siguiente expresión:

$$\sigma_A \approx \sqrt{\sum_{i=1}^{N}\left(\frac{\partial A}{\partial y_i}\right)^2 \sigma_{Yi}^2} \qquad y \qquad \sigma_B \approx \sqrt{\sum_{i=1}^{N}\left(\frac{\partial B}{\partial y_i}\right)^2 \sigma_{Yi}^2}$$

donde σ_{Yi} es la desviación estándar para cada valor Y_i .

Calculando las derivadas parciales de *B* y *A*, ecuaciones (10) y (11), y suponiendo que las σ_{Yi} son iguales para todos los Y_i ($\sigma_{Yi} = \sigma_Y \quad \forall i$) se obtienen las siguientes expresiones,

$$\sigma_A \approx \sigma_Y \sqrt{\frac{1}{N\left(\overline{X^2} - \overline{X}^2\right)}} \qquad (12) \qquad\qquad \sigma_B \approx \sigma_Y \sqrt{\frac{\overline{X^2}}{N\left(\overline{X^2} - \overline{X}^2\right)}} \qquad (13)$$

Donde se considera que las desviaciones de las *N* mediciones de las Y_{ik} alrededor de su promedio Y_i son del mismo orden que las diferencias e_i

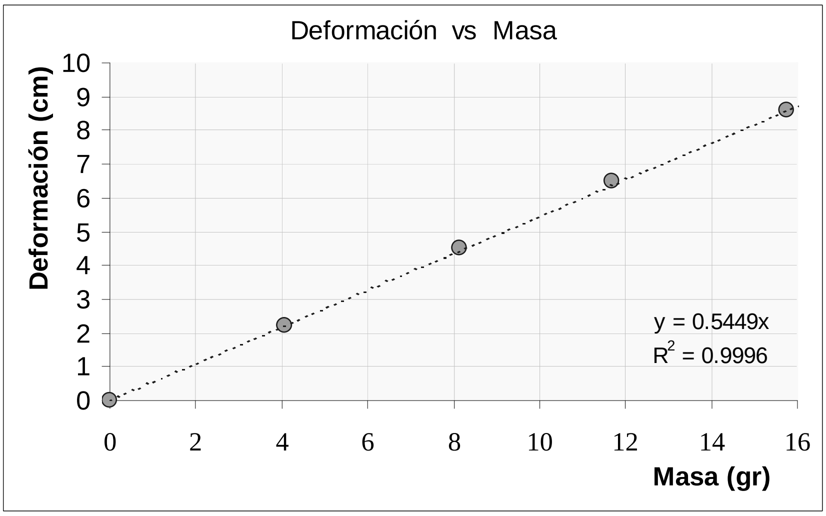
$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N-2}}$$

► *¿Cómo se evalúa la bondad del ajuste?*

Una vez que tenemos la recta de regresión, es necesario saber si el ajuste que ofrece sobre la nube de puntos es suficientemente bueno. La medida más comúnmente utilizada para evaluar qué tan bueno es el ajuste de la recta de regresión es el **coeficiente de correlación lineal *R***, que se define:

$$R = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sqrt{\left(\overline{X^2} - \overline{X}^2\right)\left(\overline{Y^2} - \overline{Y}^2\right)}}$$

Los valores *R* pueden ir de + 1 a - 1, pasando por el cero. El coeficiente de correlación nos informa acerca del grado de relación (lineal) entre dos variables. Si la relación es lineal perfecta, *R* será 1 ó -1. El coeficiente *R* será positivo si la relación es positiva (al aumentar *X* aumenta *Y*), y *R* será negativo en el caso contrario (si al aumentar *X*, disminuye *Y*). Si $R \approx 0$ puede indicar dos cosas: que las variables no están correlacionadas o que el ajuste a un modelo lineal no es el adecuado.



En la figura se muestra el resultado de aplicar el ajuste de mínimos cuadrados a los valores medidos de deformación de un resorte en función de la masa del cuerpo que pende del mismo. Los puntos son los datos experimentales y la línea es la recta de ajuste.

A un costado del gráfico se presenta la ecuación de la recta de ajuste y el valor del cuadrado del coeficiente de correlación.

► ***¿En qué casos se puede aplicar este método?***

Es importante notar que el método de regresión lineal puede también aplicarse a relaciones no lineales, como por ejemplo la función potencial $Y = B\,X^A$. Esta forma funcional es muy común en las ciencias ya que, con frecuencia, la curva que representa se aproxima al comportamiento de gran variedad de sistemas. Supongamos que los datos medidos (X_i,Y_i) están relacionados entre sí por una curva de este tipo, si se representa Y en función de X lo que se tiene (en el caso de $A\neq1$) es una curva cóncava hacia arriba ($A>1$) o hacia abajo ($A<1$).

Para poder emplear el análisis de regresión lineal, bastará con transformar la función potencial en una relación lineal. Esto se logra aplicando en ambos miembros de la igualdad la función logaritmo

$$\log Y = \log(B.X^A) = \log B + A \log X$$

y realizando un cambio de variables $Y^*=\log Y$ y $X^*=\log X$. De esta forma se obtiene una relación lineal $Y^* = \log B + AX^*$ donde A es el valor de la pendiente y $\log B$ la ordenada al origen.

Ejemplo: Se mide el período T de un péndulo simple para distintas longitudes de L. En el caso de pequeñas amplitudes de oscilación, ambas variables están relacionadas por $\mathrm{T} = 2\pi\sqrt{L/g}$ donde g es la aceleración de la gravedad. La relación es del tipo $\mathrm{T} = aL^b$ con $a = 2\pi/g$ y $b = 1/2$. Mediante el ajuste por regresión lineal de T en función de L^b se puede obtener el valor de la constante a e indirectamente del valor de la aceleración de la gravedad g.

Bibliografía:

1. *Introducción a las mediciones de laboratorio.* A. Maiztegui y R. Gleiser. Ed. Kapeluz – 1980 (En Biblioteca Central)

2. *An introduction to error analysis.* John R. Taylor (1997) University Science Books (disponible como apunte digital en la cátedra)

3. *Mecánica Elemental.* Juan. G. Roederer. Eudeba (1963)

4. *Física Re-Creativa.* Salvador Gil y Eduardo Rodriguez. Prentice Hall (2001), disponible en http://www.fisicarecreativa.com/libro/indice_exp.htm#metrologia