

Nociones sobre cuadrados mínimos

1. Introducción

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:

$$\{x_i \pm u(x_i), y_i \pm u(y_i)\}; i = 1 \text{ hasta } n$$

donde $u(x_i)$, $u(y_i)$ son las incertidumbres de los datos. Supongamos además que hay motivos suficientes como para conjeturar que existe una relación funcional lineal entre X e Y (si la relación funcional es más compleja, lo que sigue puede aún aplicarse de varias maneras), es decir:

$$Y = a + bX$$

En este caso, cabe plantearse un doble problema: ¿me permiten mis datos afirmar que la relación lineal es correcta (al menos dentro del rango de valores medidos)? Y, supuesto que esto tiene respuesta afirmativa ¿cómo determino los parámetros a y b , y cuales son sus incertezas?

La técnica más usual para responder a esas preguntas, desarrollada originalmente por Gauss, se conoce como el método de ajuste por cuadrados mínimos, y se basa en lo que sigue.

Sean a y b los valores que se quieren encontrar. Para cada par de datos x_i, y_i , se define la cantidad: $d_i = y_i - (a + bx_i)$. Es intuitivo ver que, si la relación es efectivamente lineal, y los a y b son los correctos, cada d_i debe ser “pequeño” (luego precisaremos mejor esto). No podemos esperar que sean cero, porque hay incertezas de medición tanto en X como en Y , pero lo razonable es esperar que, si los valores de a y b elegidos son los que corresponden a la recta buscada, todos los d_i sean chicos. Como criterio que satisface esto, se puede proponer que la suma de los d_i sea mínima (en realidad, esto da el criterio para evaluar a y b). Sin embargo, la suma simple no es buen criterio, porque los d_i pueden ser positivos y negativos, y por tanto cancelarse mutuamente haciendo que la suma sea mínima aunque los d_i sean grandes en módulo. Para evitar este problema, se pueden sumar los módulos o, lo que es más conveniente, los d_i elevados al cuadrado. Entonces, la primera aproximación del método consiste en elegir los valores de a y b que minimizan a la suma:

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

2. Cuadrados mínimos ponderados

Antes de ver como se obtienen a y b usando ese proceso de minimización, vamos a dar un paso más para decidir que suma minimizar. Para ello, vamos a suponer (de momento arbitrariamente, luego se precisará esto) que las incertidumbres de la magnitud X son despreciables frente a las incertidumbres de Y . Nótese, de paso, que d_i es la distancia, en la dirección del eje y , entre y_i y el valor $y'_i (= a + bx_i)$ que es la coordenada y del punto de la recta buscada cuya abscisa es x_i . Es claro que, si los x_i “no tienen incertidumbre”, es cierto que la relación es lineal y que sus parámetros son

a y b , entonces las diferencias entre y_i e y'_i son únicamente atribuibles a las incertidumbres en los y_i . Pero si estas incertidumbres son distintas entre sí, parece sensato realizar el procedimiento de minimización asignando una mayor importancia a los d_i provenientes de valores de los y_i que tengan incertidumbres más chicas. 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 $u(y_i)$. La forma más sencilla de hacer esto es dividir a cada d_i por $u(y_i)$: cuanto menor es la incertidumbre, mayor es el coeficiente que tiene el d_i . Usando este criterio, lo que se debe minimizar es:

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

Al método de obtención de a y b usando la minimización de Σ' se lo llama “cuadrados mínimos ordinarios”, mientras que al que minimiza a Σ se lo llama “cuadrados mínimos ponderados”. Como la técnica de minimización es común, describiremos sólo la de cuadrados mínimos ponderados.

Una parte considerable del problema original que tenemos planteado consiste en responder a esta pregunta: dados n valores de x_i, y_i , y $u(y_i)$ ¿cuales son los valores de a y b que minimizan a Σ ? El análisis matemático enseña que son la solución del siguiente sistema de ecuaciones:

$$\frac{\partial \Sigma}{\partial a} = 0; \frac{\partial \Sigma}{\partial b} = 0$$

Haciendo esas derivadas se obtiene:

$$\begin{aligned} \frac{\partial \Sigma}{\partial a} &= \frac{\partial}{\partial a} \sum_i \left(\frac{y_i - a - bx_i}{u(y_i)} \right)^2 = \sum_i \frac{\partial}{\partial a} \left(\frac{y_i - a - bx_i}{u(y_i)} \right)^2 = -2 \sum_i \left(\frac{y_i - a - bx_i}{u(y_i)^2} \right) = 0 \\ \frac{\partial \Sigma}{\partial b} &= \frac{\partial}{\partial b} \sum_i \left(\frac{y_i - a - bx_i}{u(y_i)} \right)^2 = -2 \sum_i x_i \left(\frac{y_i - a - bx_i}{u(y_i)^2} \right) = 0 \end{aligned}$$

Desarrollando las sumatorias, y definiendo (para simplificar notación):

$$\begin{aligned} S &= \sum \frac{1}{u(y_i)^2}; S_x = \sum \frac{x_i}{u(y_i)^2}; S_y = \sum \frac{y_i}{u(y_i)^2} \\ S_{xx} &= \sum \frac{x_i^2}{u(y_i)^2}; S_{yy} = \sum \frac{y_i^2}{u(y_i)^2}; S_{xy} = \sum \frac{x_i y_i}{u(y_i)^2} \end{aligned}$$

el sistema de ecuaciones se escribe como:

$$\begin{aligned} S_y - aS - bS_x &= 0 \\ S_{xy} - aS_x - bS_{xx} &= 0 \end{aligned}$$

cuya solución es:

$$b = \frac{SS_{xy} - S_x S_y}{SS_{xx} - S_x^2}; u(b)^2 = \frac{S}{SS_{xx} - S_x^2}; a = \frac{S_{xx} S_y - S_x S_y^2}{SS_{xx} - S_x^2}; u(a)^2 = \frac{S_{xx}}{SS_{xx} - S_x^2}$$

También se han agregado expresiones para las incertidumbres de a y b . No se dará aquí el detalle de su cálculo, basta comentar que se obtienen de las expresiones de a y b , propagando las incertidumbres $u(y_i)$, que afectan a cada y_i que aparece en las

sumatorias ($u(a)^2 = \sum_j \left(\frac{\partial a}{\partial y_j} u(y_j) \right)^2$) y otra equivalente para $u(b)^2$.

3. Confiabilidad

Queda todavía por contestar a la pregunta de si los datos permiten confiablemente asumir que las variables X y Y guardan una relación lineal. Este tema está un poco confuso en la literatura, y es frecuente encontrar que el criterio para sostener la relación lineal es que un coeficiente, r^2 , llamado coeficiente de correlación de los datos y definido como sigue:

$$r^2 = \frac{(SS_{xy} - S_x S_y)^2}{(SS_{xx} - S_x^2)(SS_{yy} - S_y^2)}$$

sea “suficientemente cercano a 1”. Haciendo un poco de maniobras algebraicas, se puede demostrar que si $r^2=1$, entonces todos los puntos (x_i, y_i) están alineados en una recta. Obviamente, si se cumple esa condición, la relación lineal queda asegurada.

En la interpretación del coeficiente de correlación se debe tener en cuenta que:

- $r = \pm 1$ indica una relación lineal exacta positiva (creciente) o negativa (decreciente),
- $r = 0$ indica la no existencia de relación lineal estocástica, pero no indica independencia de las variables ya que puede existir una relación no lineal incluso exacta,
- valores intermedios de r ($0 < r < 1$ ó $-1 < r < 0$) indican la existencia de una relación lineal estocástica, más fuerte cuanto más próximo a $+1$ (ó -1) sea el valor de r .

Ya se mencionó que, gráficamente, d_i es la distancia en la dirección del eje y entre el dato i -ésimo y la recta obtenida usando el método de cuadrados mínimos (es decir, aquella determinada por los valores de a y b que se obtienen del método). Por lo tanto, $d_i/u(y_i)$ es una “medida” de cuanto vale esa distancia comparada con el error en Y en ese punto. Supongamos que las incertidumbres en Y son de naturaleza instrumental (las últimas recomendaciones de Comités Internacionales sugieren no hacer esta discriminación y tratar a todas las incertidumbres como estadísticas, pero mantendremos aca la diferenciación). Es claro que, si ese cociente es menor o igual que 1, para ese dato particular, se puede decir que el punto pertenece (experimentalmente) a la recta. Si las incertidumbres fueran estadísticas gaussianos, esta condición no es suficiente, porque hay una probabilidad respetable (33%) de que el valor medio utilizado para y_i caiga fuera del intervalo definido por una dispersión, y en estos casos se debería agregar un intervalo de probabilidad en esa condición. Dejando de lado este problema, consideremos ahora que pasa con el conjunto de datos. En principio, es posible que algunos de los puntos tengan ese cociente menor

que 1 mientras que otros lo tengan mayor que 1. Parece razonable usar como criterio para decidir si se acepta o no la relación lineal el que la mayoría de los puntos satisfaga esa condición. Y esto se puede hacer sencillamente calculando el valor que toma Σ , una vez calculados a y b . En efecto, Σ es la suma de los cuadrados de esos cocientes, de modo que Σ/n es el valor medio de ese cuadrado. Luego, $(\Sigma/n)^{1/2}$ es una estimación de cuanto vale el promedio de los $d_i/u(y_i)$, y si éste resulta ser menor o igual que 1, la relación lineal debe ser aceptada; si es mucho mayor que 1, debe ser rechazada. Entre el valor 1 y “mucho mayor que 1” queda todavía un margen de indecisión (entre otras cosas, en decidir cuanto es “mucho”). La ventaja de la teoría estadística mencionada más arriba consiste en que, haciendo una comparación que es substancialmente equivalente, da resultados en términos de probabilidades para cualquier valor que se obtenga en la comparación. Con todo, el criterio expuesto aquí tiene la ventaja de ser sencillo de entender y válido para cualquier tipo de incertidumbres.

¿Que relación tiene esto con el r^2 ? Si se reemplazan los resultados de a y b dados más arriba en la expresión de Σ , y se usa la definición de r^2 , haciendo un poco de maniobras algebraicas se encuentra que:

$$\Sigma = (1 - r^2)(S_{yy} - S_y^2/S)$$

en donde se ve que si r^2 vale 1, Σ es cero y por tanto, el criterio dado anteriormente se satisface. Sin embargo, el hecho que r^2 sea cercano a 1 (=0,95, por ejemplo) no alcanza para garantizar que Σ sea pequeño (menor o del orden de n) porque el otro factor $(S_{yy} - S_y^2/S)$ es habitualmente un número mucho mayor que 1.

Todavía no se dijo nada sobre n , el número de datos. Si los datos fueran sólo 2, es obvio que determinan un único valor de pendiente y ordenada al origen, y consecuentemente, $\Sigma=0$. Pero también es claro que, con sólo dos datos, no se puede pretender asegurar una relación lineal entre las variables X e Y . También es claro que, cuantos más datos se tengan, mayor confianza se puede tener en determinar la existencia (o inexistencia) de una relación lineal. En un tratamiento estadístico existen criterios adicionales sobre el número de datos requeridos, pero para lo que nos interesa acá basta decir que un mínimo razonable es $n = 10$. De todos modos, hay que agregar otra consideración: no basta con tener un valor de n razonable; es necesario además que los rangos de valores de X e Y cubiertos por los datos sean también razonables, al menos que cubran un orden de magnitud. Nótese que, en un rango de valores relativamente pequeño, cualquier función puede ser bien aproximada por una recta.