Métodos de mínimos cuadrados

1. Introducción

El método de mínimos cuadrados tiene una larga historia que se remonta a los principios del siglo XIX. En Junio de 1801, Zach, un astrónomo que Gauss había conocido dos años antes, publicaba las posiciones orbitales del cuerpo celeste Ceres, un nuevo "pequeño planeta" descubierto por el astrónomo italiano G. Piazzi en ese mismo año. Desafortunadamente, Piazzi sólo había podido observar 9 grados de su órbita antes de que este cuerpo desapareciese tras del sol. Zach publicó varias predicciones de su posición incluyendo una de Gauss que difería notablemente de las demás. Cuando Ceres fue redescubierto por Zach en Diciembre de 1801 estaba casi exactamente en donde Gauss había predicho. Aunque todavía no había revelado su método, Gauss había descubierto el método de mínimos cuadrados. En un trabajo brillante logró calcular la órbita de Ceres a partir de un número reducido de observaciones, de hecho, el método de Gauss requiere sólo un mínimo de 3 observaciones y todavía es, en esencia, el utilizado en la actualidad para calcular las órbitas. Uno de los recursos más usados por los investigadores para evaluar el grado de acuerdo entre los resultados de un experimento y una teoría o modelo consiste en representar ambos conjuntos de datos, en un mismo gráfico. Existe un generalizado consenso en la literatura científica en utilizar símbolos para representar los resultados de un experimento y líneas continúas para describir las expectativas teóricas. Con frecuencia se encuentran situaciones en las que los datos (x_i, y_i) sugieren una relación lineal entre dos variables, es decir una expresión: $y = a.x + b$. La pregunta que deseamos responder es: ¿cuáles son los parámetros **a** y **b** de la recta que mejor se ajustan a los datos? El método de cuadrados mínimos es un procedimiento que permite responder esta pregunta. Cuando la relación entre las variables x e y es lineal, el método de ajuste por cuadrados mínimos se denomina también método de regresión lineal.

2. Método de mínimos cuadrados

Supongamos que tenemos un conjunto de N mediciones x_i, y_i, donde supondremos que x_i es la variable que controlamos experimentalmente, *variable independiente*. A la variable y_i, vinculada con x_i a través del modelo, la denominaremos *variable dependiente*. Supondremos, y esto es una hipótesis muy fuerte, que la desviación estándar de la variable independiente puede ser considerada nula, es decir la dispersión en torno al valor esperado de la variable x_i tiende a cero, $\sigma_{x_i} = 0$. Para la variable y_i la desviación estándar es σ_{y_i}.

En la figura 1, se observa que los datos experimentales tienen una tendencia aproximadamente lineal. Nuestro objetivo es encontrar la mejor recta que los ajusta, o sea los valores de $\boldsymbol{a}$ y $\boldsymbol{b}$ de la recta:

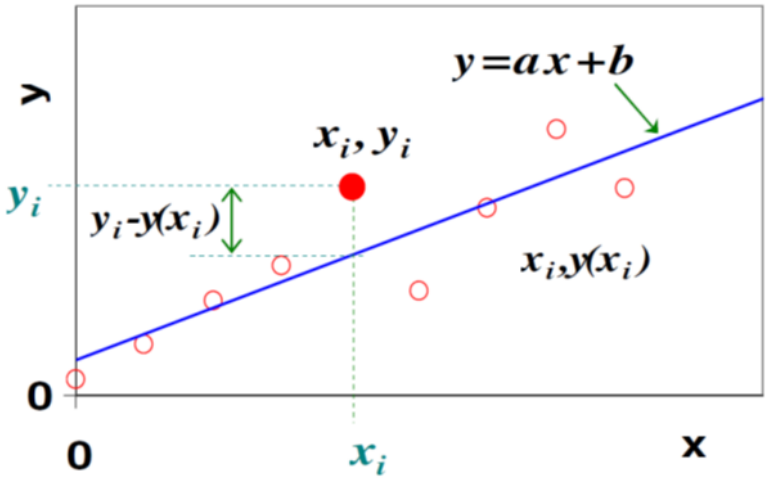

Figura 1: Representación gráfica de (x_i, y_i) con tendencia lineal. Los círculos representas valores observados. La recta es la representación del modelo $y(x) = ax + b$. La cantidad $y_i - y(x_i)$ es la desviación de cada observación de y_i respecto del valor predicho por el modelo $y(x_i)$.

$$y = ax + b, \tag{1}$$

que mejor describen los datos observados. Para ello resulta útil definir la función χ^2

$$\chi^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 \tag{2}$$

Esta función,χ^2, es una medida de la desviación total al cuadrado, $[y_i - y\,(x_i)]^2$, de los valores observados yi respecto de los predichos por el modelo lineal a.x+b. En otras palabras, χ^2 es una medida de la distancia (vertical) de todos los datos (x_i, y_i), a la recta. Para un dado conjunto de datos (x_i, y_i), el valor deχ^2 depende de los parámetros de la recta, a y b. El método de cuadrados mínimos supone que los valores de la pendiente a y la ordenada al origen b, que mejor ajustan los datos, son aquellos que minimizan esta desviación total, o sea, los que minimizan la función $\chi^2(a, b)$. El problema de minimización se reduce al de resolver el par de ecuaciones:

$$\frac{\partial \chi^2}{\partial a} = 0 \qquad \text{y} \qquad \frac{\partial \chi^2}{\partial b} = 0 \tag{3}$$

cuyas incógnitas son **a** y **b**.

$$\frac{\partial \chi^2(a,b)}{\partial a} = 2a \sum x_i^2 + 2b \sum x_i - 2 \sum x_i y_i = 0$$

$$\frac{\partial \chi^2(a,b)}{\partial b} = 2bN + 2a \sum x_i - 2 \sum y_i = 0$$

Resolviendo estas ecuaciones, resulta

$$a = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \tag{4}$$

y

$$b = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \tag{5}$$

La recta obtenida con estos coeficientes se denomina *línea de regresión*. Los resultados 4 y 5 se aplican cuando todos los datos de la variable dependiente tienen la misma incertidumbre absoluta y la incertidumbre de la variable independiente se considera despreciable. Una medida de la calidad o bondad del ajuste realizado viene dada por el *coeficiente de correlación de Pearson* R^2 entre las variables x e y, que adopta valores entre 0 y 1 y caracteriza la dispersión de los datos alrededor de la línea de cuadrados mínimos. Consideremos las desviaciones de los puntos observados, (x_i, y_i): A) respecto de la recta obtenida de cuadrados mínimo y B) respecto a la recta horizontal $y = \bar{y}$, donde $\bar{y}$ es el promedio de los valores y_i. Si la recta de cuadrados mínimos es una buena descripción de los datos, los valores (x_i, y_i) se agrupan a lo largo de la línea de regresión. La suma de los cuadrados de las desviaciones a esta línea, representados por χ^2, debería ser menor que la suma de los cuadrados de las desviaciones a la línea horizontal $y = \bar{y}$. Se define el coeficiente de correlación al cuadrado, R^2:

$$R^2 = \frac{\sum (y_i - \bar{y})^2 - \sum \left[y_i - (ax_i + b)\right]^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (y_i - \bar{y})^2 - \chi^2}{\sum (y_i - \bar{y})^2} \tag{6}$$

El primer término en el numerador es la suma de los cuadrados de las desviaciones de los puntos de la línea horizontal que pasa por y . El segundo término es la suma de los cuadrados de las desviaciones de los puntos de la línea de regresión $y = ax+b$, o sea por χ^2 definido por la Ec.2. Nótese que R^2 es adimensional. Si los datos caen exactamente sobre la línea de regresión, hay correlación perfecta, el segundo término es aproximadamente cero ($\chi^2 \approx 0$) y $R^2 \approx 1$. Por otro lado, a medida que peor es el ajuste, mayor será el valor de χ^2. El valor máximo que puede alcanzar χ^2 es del orden de$\sum (y_i - \bar{y})^2$, en este caso, no hay correlación entre las variables x e y, y el numerador de la Ec.(7.19) es cero, o sea:$R^2 \approx 0$. Cuando en 6 se sustituyen las

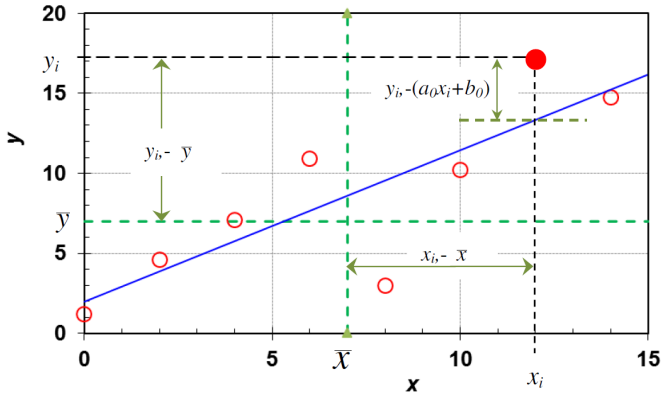

Figura 2: Datos empíricos (círculos) que se agrupan a lo largo de una recta $y = ax+b$. Las desviaciones de los puntos de la recta de cuadrados mínimos y las desviaciones de los puntos a la recta horizontal sirven para definir al coeficiente R^2 [Ec.6]

ecuaciones 4 y 5 para a y b, se obtiene para R^2:

$$R^2 = \left[\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N\sigma_x\sigma_y}\right]^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \cdot \sigma_y^2} \tag{7}$$

donde $\sigma_x^2 \equiv < x^2 > - < x >^2$, $\sigma_y^2 \equiv < y^2 > - < y >^2$ y $\sigma_{xy} = [\sum (x_i - \bar{x})(y_i - \bar{y})] / N$

Si $R^2 \approx 1$, decimos que el modelo lineal es adecuado para describir los datos experimentales y hay buena correlación (lineal) en los datos de x e y. Cuando $R^2 \approx 0$, decimos que la expresión lineal no es una descripción adecuada de los datos. En este caso, conviene analizar detenidamente el gráfico y buscar si hay alguna relación no–lineal que aproxime mejor la dependencia de x con y. Si $R^2 \approx 0$ puede indicar que no hay ninguna correlación entre las variables. Sin embargo, un valor de $R^2 \approx 0$ no implica necesariamente que no haya correlación entre las variables, sólo significa que la relación lineal entre ellas no es adecuada. Así, si los pares de puntos (x, y) describen una circunferencia, tenderemos que $R^2 \approx 0$, ver figura3d). Desde luego, si los pares (x, y) no tienen correlación alguna entre ellos, también tendríamos $R^2 \approx 0$ como se ilustra en la figura3c).

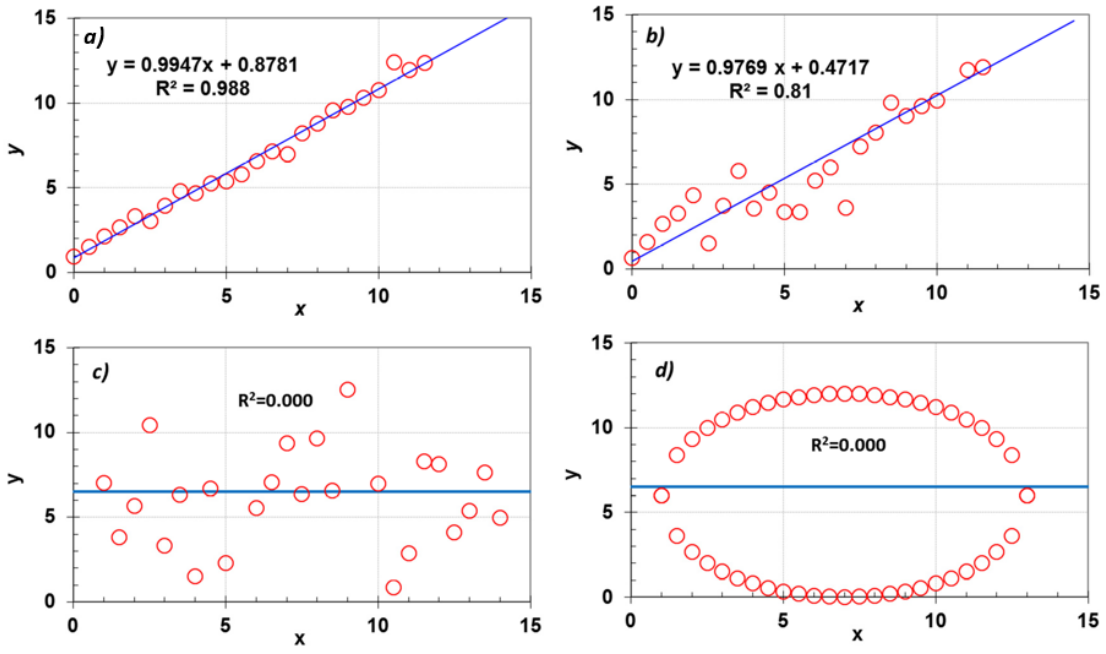

Figura 3: Ajuste de datos experimentales por un modelo lineal. a) Caso de una buena correlación lineal, b) aceptable, c) es un caso en el que prácticamente no hay correlación entre x e y, d) existe una buena correlación pero el modelo lineal es inadecuado.

3. Casualidad, correlación

Una observación importante a tener en cuenta es que una correlación entre dos conjuntos de datos x e y, no siempre implica una relación de causalidad entre ellos. En otras palabras, si $R^2 \approx 1$, esto no significa necesariamente que y depende causal-

mente de x o viceversa. La correlación entre las variables en una condición necesaria pero no suficiente para exista una dependencia causal entre ellas. Es una falacia atribuir causalidad a dos eventos que ocurren a la vez, es decir afirmar que "*Si aparecen juntos es que son causa y efecto*". Un ejemplo de lo anterior fue la observación que "*Los niños que duermen con la luz encendida son más propensos a desarrollar miopía en la edad adulta.*" Este estudio fue realizado en un centro médico de la Universidad de Pensilvania y llegó a la revista Nature en mayo de 1999. Estudios posteriores encontraron una falacia en esta conclusión. Observaciones más cuidadosas muestran un importante carácter hereditario en la miopía. Como los padres miopes requieren de buena iluminación para ver, ellos tienden a dejar más luces encendidas en las habitaciones de sus hijos.

4. Incertidumbre de los coeficientes

En muchos casos, el objetivo de un determinado estudio experimental consiste obtener los parámetros de un ajuste. Por ejemplo, si deseamos determinar la constante elástica k de un resorte a partir de mediciones de las fuerzas aplicadas F_i y sus respectivos estiramientos x_i, suponiendo que existe una relación lineal entre estas variables, del tipo $F = kx$. En este caso, el valor de k será precisamente la pendiente de la recta que mejor se ajusta a los datos de F_i como función de x_i. La pregunta que queremos responder ahora, es cuales son las incertezas en estos parámetros obtenidos por cuadrados mínimos. Resulta útil disponer de un modo de estimar las incertidumbres asociadas a la determinación de los parámetros a y b de las Ecs. 4 y 5, que denotaremos con los símbolos σ_a y σ_b. Sólo presentamos los resultados. Las incertidumbres de los parámetros del ajuste vienen dadas por las expresiones

$$\sigma_a = \sqrt{\frac{\chi^2_N}{N\,Var(x)}} \tag{8}$$

$$\sigma_b = \sqrt{\frac{\chi^2_N \sum_1^N x_i^2}{N\,Var(x)}} = \sigma_a \sqrt{\langle x^2 \rangle} \tag{9}$$

donde χ^2_N, conocido como el valor de Chi-cuadrado por grado de libertad, viene dada por:

$$\chi^2_N = \frac{1}{N-2}\chi^2 \tag{10}$$

y

$$Var(x) = \frac{\sum_1^N x_i^2}{N} - \left[\frac{\sum_1^N x_i}{N}\right]^2 = \langle x^2 \rangle - \langle x \rangle^2 \tag{11}$$

o en función de R^2 queda:

$$\sigma_a = a\sqrt{\frac{1}{N-2}\left(\frac{1}{R^2} - 1\right)} \qquad \text{y} \qquad \sigma_b = \sigma_a \sqrt{\langle x^2 \rangle} \tag{12}$$

con $\langle x^2 \rangle = \sum x_i^2 / N$