Métodos cuantitativos de análisis gráfico

Método de cuadrados mínimos – Regresión lineal

Hemos enfatizado sobre la importancia de las representaciones gráficas y hemos visto la utilidad de las versiones linealizadas de los gráficos (*X*, *Y*) junto a las distintas maneras de llevar a cabo la linealización. A menudo nos confrontamos con situaciones en las que existe o suponemos que existe una relación lineal entre las variables *X* e *Y*. Surge de modo natural la pregunta: ¿cuál es la relación analítica que mejor se ajusta a nuestros datos? El *método de cuadrados mínimos* es un procedimiento general que nos 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*. En este capítulo discutiremos este último caso. El lector puede consultar en el Apéndice F de la Ref. [1] una discusión del caso general de cuadrados mínimos cuando el modelo es no lineal y los datos están afectados de errores.

La Fig. 1 ilustra el caso lineal. La dispersión de los valores está asociada a la fluctuación de los valores de cada variable. Observamos o suponemos una tendencia lineal entre las variables y nos preguntamos sobre cuál es la *mejor recta*:

$$y(x) = a\,x + b \tag{1}$$

que representa este caso de interés.

Es útil definir la función χ^2 (Chi-cuadrado)[1-3]:

$$\chi^2 = \Sigma_i \left(y_i - (a \cdot x_i + b) \right)^2 \tag{2}$$

que es una medida de la desviación total de los valores observados y_i respecto de los predichos por el modelo lineal $a\,x + b$. Los mejores valores de la pendiente *a* y la ordenada al origen *b* son aquellos que minimizan esta desviación total, o sea, son los valores que remplazados en la Ec.(1) minimizan la función χ^2, Ec.(2). Los parámetros *a* y *b* pueden obtenerse usando técnicas matemáticas que hacen uso del cálculo

diferencial. Aplicando estas técnicas, el problema de minimización se reduce al de resolver el par de ecuaciones:

$$\frac{d\chi^2}{da} = 0 \quad \text{y} \quad \frac{d\chi^2}{db} = 0 \tag{3}$$

de donde resulta:[1-4]

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

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

Actualmente, la mayoría de los programas de análisis de datos y planillas de cálculo, realizan el proceso de minimización en forma automática y dan los resultados de los mejores valores de *a* y *b*, o sea los valores indicado por la ecuaciones (4) y (5).

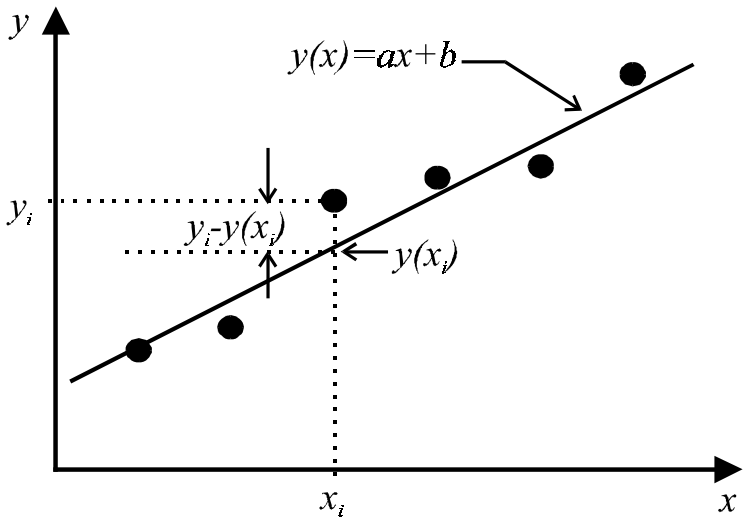

Figura 1 Gráfico de datos asociados a un modelo lineal. La cantidad y_i - $y(x_i)$ representa la desviación de cada observación de y_i respecto del valor predicho por el modelo $y(x)$.

El criterio de mínimos cuadrados reemplaza el juicio personal de quien mire los gráficos y defina cuál es la mejor recta. En los programas como Excel, Origin, etc., este cálculo se realiza usando la herramienta “regresión lineal” o “ajuste lineal”. Los resultados (4) y (5) se aplican en el caso lineal 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 dado por el *coeficiente de correlación* R^2 entre las variables X e Y, definido como:

$$R^2 = \frac{Cov(x,y)^2}{Var(x)\cdot Var(y)} \tag{6}$$

donde

$$Cov(x,y) = \frac{N\sum_{i=1}^{N} x_i \cdot y_i - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i}{N^2} \tag{7}$$

$$Var(x) = \frac{\sum_{i=1}^{N} x_i^2}{N} - \left(\frac{\sum_{i=1}^{N} x_i}{N} \right)^2 = < x^2 > - < x >^2 \tag{8}$$

y

$$Var(y) = \frac{\sum_{i=1}^{N} y_i^2}{N} - \left(\frac{\sum_{i=1}^{N} y_i}{N} \right)^2 = < y^2 > - < y >^2 \tag{9}$$

El valor de R varía entre –1 y 1. Si R es ±1 o próximo a estos valores, decimos que el modelo lineal es adecuado para describir los datos experimentales. Cuando R se aparta de estos extremos decimos que una expresión lineal no es una buena descripción de los datos. En este caso, conviene analizar el gráfico y buscar una relación no–lineal que aproxime mejor la dependencia. Dado que R mide el grado de correlación lineal entre los datos, si, por ejemplo, los pares de puntos (X, Y) tienen una relación tal que

caen sobre un círculo, aunque ellos están correlacionados, tendríamos $R \approx 0$. Desde luego, si los pares (X, Y) no tienen correlación alguna entre ellos, también tendríamos $R \approx 0$. Ver la Figura 2.

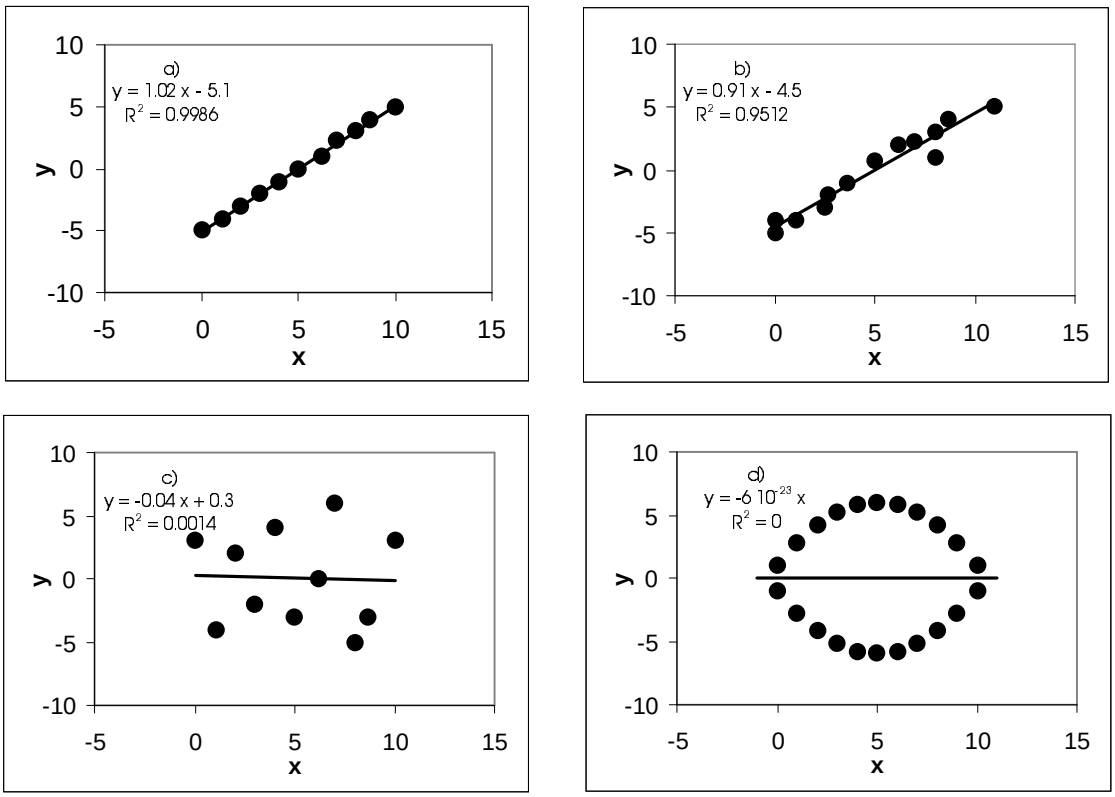

Figura 2 Ajuste de datos experimentales por un modelo lineal. a) Caso de una buena correlación lineal; b) aceptable; c) es un caso en el prácticamente no hay correlación entre X e Y; d) tiene una buena correlación pero el modelo lineal es inadecuado.

Frecuentemente el resultado que deseamos determinar de nuestro experimento es alguno de los parámetros de la Ec. (1). Por ejemplo, si deseamos determinar la constante elástica k de un resorte a partir de mediciones de fuerzas aplicadas F_i y estiramientos x_i que le producen al resorte, k será precisamente la pendiente de la recta que mejor se ajusta a los datos. Otro ejemplo es la obtención de la resistencia eléctrica R de un conductor, que deseamos determinar a partir de mediciones de tensión V_i y la corriente que lo atraviesa I_i. Por consiguiente, es útil disponer de un modo de estimar las incertidumbres asociadas a la determinación de los parámetros a y b de la Ec. (1). La importancia del método de cuadrados mínimos reside en el hecho que nos permite

obtener valores de la desviación estándar o sea los errores asociados a los parámetros *a* y *b* de la Ec. (1)[4], que denotaremos con los símbolos σ_a y σ_b. En esta sección sólo presentamos los resultados de utilidad más frecuente en el laboratorio; el lector interesado podrá encontrar un tratamiento más exhaustivo en las Ref.[1-4]. Las incertidumbres de los parámetros del ajuste vienen dadas por las expresiones:

$$\sigma_a = \sqrt{\frac{\chi_N^2}{N \cdot Var(x)}} \tag{10}$$

$$\sigma_b = \sqrt{\frac{\chi_N^2 \cdot \sum_{i=1}^{N} x_i^2}{N \cdot Var(x)}} \tag{11}$$

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

$$\chi_N^2 = \frac{1}{N-2} \cdot \chi^2 \tag{12}$$

Las incertidumbres de los parámetros *a* y *b* también pueden escribirse en términos del coeficiente de correlación *R* del siguiente modo:

$$\sigma_a = \sqrt{\frac{a^2}{(N-2)} \cdot \left(\frac{1}{R^2} - 1 \right)} \tag{13}$$

$$\sigma_b = \sigma_a \sqrt{< x^2 >} \tag{14}$$

donde

$$< x^2 >= \frac{\sum_{i=1} x_i^2}{N} \tag{15}$$

Estas expresiones son de mucha utilidad para estimar σ_a y σ_b, ya que la mayoría de las planillas de cálculo y programas de ajuste, por los regular indican los valores de los parámetros *a* y *b* que mejor ajustan los datos y el valor de *R*.

Precauciones en el análisis

No siempre es suficiente admitir que dos variables siguen una relación lineal guiándonos por lo que muestra un gráfico de los datos en escalas lineales. Menos aun si *sólo* evaluamos el coeficiente de correlación del ajuste lineal que propondríamos a partir de este gráfico. Un gráfico de $Y = X^{1.1}$ (variables sin correlación lineal) puede ajustarse por una recta y obtenerse a la vez un coeficiente de correlación lineal (inexistente) de, por ejemplo, 0.998. Un gráfico de datos experimentales de $Y = X$ con algo de dispersión fortuita de los puntos, podría devenir en un coeficiente de, por ejemplo, 0.995, menor que el anterior. Entre los coeficientes hay una diferencia, apenas, del 3 por mil. Pero en un gráfico log-log, la diferencia de pendientes será la que hay entre 1.1 y 1.0, lo que representa un 10% de discrepancia entre los exponentes de la variable *X*. Estos métodos de análisis nos enseñan que los efectos de correlación pueden estar enmascarados por el efecto del "ruido" de los datos. En ocasiones lo difícil es establecer si existe correlación entre las variables, aun cuando los datos provengan de fuentes "limpias" que hayan producido datos con relativamente poca dispersión. Muchas veces la decisión entre dos alternativas debe hacerse usando otros criterios. Por ejemplo, la consistencia con otros conjuntos de datos o sobre la base de consideraciones de simetría o concordancia con teorías bien establecidas.

Ejemplo: Imaginemos un experimento donde se mide la distancia que recorre un móvil sobre una línea recta mientras una fuerza constante actúa sobre él. Esperamos que el movimiento sea uniformemente acelerado. Supongamos que el cuerpo parte del reposo, que medimos *x(t)* a tiempos largos y que los datos colectados son los representados en la Fig. 3.a.

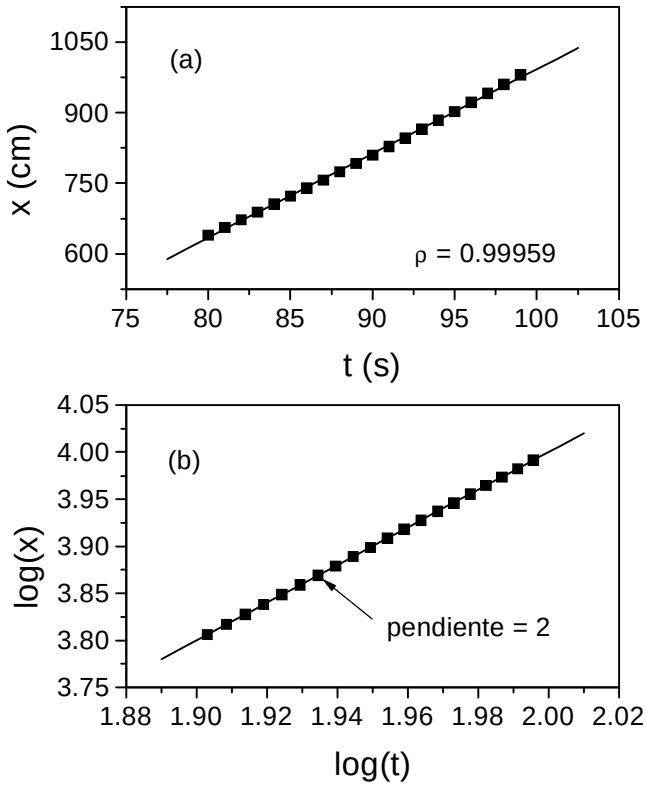

Figura 3 Representación de *x(t)* para un cuerpo que se mueve con movimiento uniformemente acelerado. (a) A tiempos largos no se aprecia bien la curvatura de la curva y, dado que el coeficiente de correlación lineal es muy cercano a la unidad, podría suponerse que la correlación es lineal. (b) *log(x)* en función de *log(t)*, de donde se deduce que la relación no es lineal sino cuadrática.

Si los datos experimentales se analizan sobre el gráfico de escalas lineales, el ajuste por un modelo lineal es más que tentador. Hecho esto, se obtiene la ecuación de la mejor recta y un coeficiente de correlación muy alto, $R = 0.99959$. Sin embargo, un modelo basado en las ecuaciones de la dinámica dice que

$$x = \frac{1}{2} a t^2$$

donde *a* es la aceleración. En la Fig. 3.b están los logaritmos de los mismos datos, de donde se ve claramente la proporcionalidad $x \propto t^2$ que predice el modelo, difícilmente demostrable a partir del gráfico de la Fig. 3.a. Evidentemente, el uso de una aproximación lineal no es buena en este problema y el mero juicio del valor del coeficiente de correlación no es suficiente.

Referencias

1. S. Gil y E.Rodríguez, *Física re-Creativa*, Prentice Hall, Buenos Aires 2001.
2. P. Bevington and D. K. Robinson, *Data reduction and error analysis for the physical sciences*, 2nd ed., McGraw Hill, New York, 1993.
3. Stuardt L. Meyer, *Data analysis for scientists and engineers*, John Willey & Sons, Inc., New York, 1975.
4. D. C. Baird, *Experimentación*, 2ª ed., Prentice Hall Hispanoamericana S.A., México, 1991.