# AJUSTE DE CURVAS (I) : REGRESIÓN LINEAL [1].

Dados n pares de datos ($x_i$, $y_i$), queremos hallar una ecuación para la "mejor" curva para este conjunto de datos. Si los datos están relacionados linealmente, entonces el proceso de llama **regresión lineal**. En general, los datos no están vinculados linealmente y el proceso de obtención de la ecuación para la mejor curva se llama **regresión no-lineal**. La técnica a ser usada para obtener la curva de mejor ajuste es el **método de cuadrados mínimos**.

## MÉTODO DE CUADRADOS MÍNIMOS.

Antes de considerar regresión lineal o no-lineal, usaremos el método de cuadrados mínimos para determinar el mejor estimador de una cantidad x.

Supongamos que una cantidad física se mide n veces, $x_i$, i=1,2,…,n. Un ejemplo es la medida de un período de un péndulo simple n veces, donde para cada medida la longitud, la masa y la amplitud son constantes. El método de cuadrados mínimos establece que el mejor estimador del resultado de las n medidas es aquel que minimiza la suma de los cuadrados de las desviaciones de las medidas de su mejor estimador x, esto es, minimizamos

$$\sum_{i=1}^{n}\left(x-x_i\right)^2 \qquad (1)$$

donde x es el mejor estimador desconocido. Minimizando la expresión (1) y resolviendo para x, encontramos que

$$\begin{aligned}&\frac{d}{dx}\sum_{i=1}^{n}\left(x-x_i\right)^2=0\\&2nx-2\sum_{i=1}^{n}x_i=0\\&x=\frac{1}{n}\sum_{i=1}^{n}x_i\equiv\overline{x}\end{aligned} \qquad (2)$$

Entonces, el mejor estimador es el promedio o valor medio, $\overline{x}$ .

Notar que minimizar la suma de los desvíos cuadráticos es equivalente a maximizar la probabilidad $P(x_1, x_2,\ldots, x_n)$ de obtener nuestro conjunto de medidas $x_1, x_2,\ldots, x_n$. Suponemos que los datos ($x_i$) están distribuidos de acuerdo a la distribución Gaussiana; luego, la probabilidad de obtener una medida dentro de un intervalo dx de $x_i$ es

$$P(x_i)=\frac{1}{\sigma(2\pi)^{1/2}}\exp\left[-\frac{(x-x_i)^2}{2\sigma^2}\right]dx \qquad (3)$$

donde

$$x = mejor\ estimador\ para\ x_i \tag{4}$$

y σ es el desvío estándar teórico.

La probabilidad de obtener nuestro conjunto de medidas es

$$P(x_1, x_2, ..., x_n) = P(x_1)P(x_2)...P(x_n) = \left(\frac{dx}{\sigma\sqrt{2\pi}}\right)^n \times \exp\left[-\sum_{i=1}^{n}\frac{(x-x_i)^2}{2\sigma^2}\right] \tag{5}$$

Si minimizamos el exponente en la ecuación (5), entonces $P(x_1,x_2,\ldots,x_n)$ será un máximo. La suma en el exponente se llama **suma de cuadrados mínimos**,

$$\sum_{i=1}^{n}\frac{(x-x_i)^2}{2\sigma^2} \tag{6}$$

y minimizarla es equivalente a minimizar $\sum_{i=1}^{n}(x-x_i)^2$ , puesto que σ se supone una constante.

*Nota*: Suponemos que los datos siguen la distribución de Gauss, y el método de cuadrados mínimos se usa para hallar el valor más probable.

## MÉTODO DE CUADRADOS MÍNIMOS Y REGRESIÓN LINEAL.

Dado un conjunto de n pares de puntos experimentales ($x_i$, $y_i$) ( por ejemplo, $x_i$ podría ser el tiempo y $y_i$ la velocidad promedio de un objeto que cae), nos gustaría hallar la ecuación para la mejor línea recta, como se muestra en la Figura 1. Hacemos las siguientes suposiciones:

1. Los valores medidos ($x_i$, $y_i$) están distribuidos según la distribución de Gauss (esto es así generalmente si los errores son aleatorios).
2. Los errores en $x_i$, $dx_i$, son despreciables en comparación con los errores en $y_i$, $dy_i$ (en consecuencia, solo consideramos la distribución de los valores $y_i$).
3. Los errores en y son todos iguales: $dy_1 = dy_2 = \ldots = dy_n$ (entonces, el desvío estándar $s_y$ es constante).

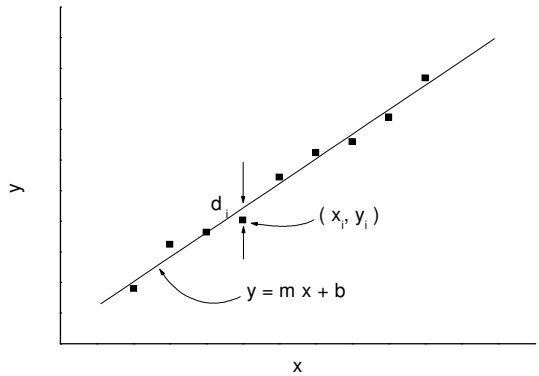


Figura 1.

La minimización de la suma de cuadrados mínimos da la ecuación para la mejor línea recta.

Aproximamos el conjunto de n medidas ($x_i$, $y_i$) por la relación lineal

$$y(x) = a_0 + a_1 x \tag{7}$$

La probabilidad de obtener el valor observado $y_i$ es

$$P(y_i) \propto \frac{1}{\sigma_y} \exp\left[ -\frac{\left[ y_i - y(x_i) \right]^2}{2\sigma_y^2} \right] \tag{8}$$

donde

$$y(x_i) = mejor\ estimador\ \ para\ \ y_i = a_0 + a_1 x_i \tag{9}$$

y $\sigma_y$ es el desvío estándar teórico. La probabilidad $P(y_1, \ldots, y_n)$ de obtener el conjunto de medidas es

$$\begin{aligned} P(y_1,...,y_n) &= P(y_1)P(y_2)...P(y_n) \\ &\propto \frac{1}{\left(\sigma_y\right)^n} \exp\left[ -\sum_{i=1}^{n} \frac{\left[ y_i - a_0 - a_1 x_i \right]^2}{2\sigma_y^2} \right] \end{aligned} \tag{10}$$

Queremos que esta probabilidad sea un máximo; entonces, el exponente (suma de cuadrados mínimos) debe ser un mínimo. Minimizando la suma de cuadrados mínimos da la ecuación para la mejor línea recta.

En la Figura 1, $d_i$ es la distancia vertical desde cada punto ($x_i$, $y_i$) a la línea $y = a_0 + a_1 x$. Deseamos hallar valores de $a_0$ y $a_1$ de manera que minimicen la función $M(a_0,a_1)$ definida como

$$M(a_0, a_1) = \sum_{i=1}^{n} \frac{d_i^2}{2\sigma_y^2} = \sum_{i=1}^{n} \frac{\left[ y_i - \left( a_0 + a_1 x_i \right) \right]^2}{2\sigma_y^2} \tag{11}$$

que es el exponente en la ecuación (10). Desarrollando el término cuadrático e ignorando la (supuesta) constante $\sigma_y$, encontramos que

$$M = \sum_i \left( y_i \right)^2 - 2a_1 \sum_i x_i y_i - 2a_0 \sum_i y_i + a_1^2 \sum_i x_i^2 + 2a_0 a_1 \sum_i x_i + n a_0^2 \tag{12}$$

A continuación hacemos

$$\frac{dM}{da_0} = 0 \quad y \qquad \frac{dM}{da_1} = 0 \tag{13}$$

para encontrar $a_0$ y $a_1$ correspondientes al mínimo valor de M. Esto produce dos ecuaciones simultáneas

$$\frac{dM}{da_0} = -2\sum_i y_i + 2a_1\sum_i x_i + 2na_0 = 0$$
$$\frac{dM}{da_1} = -2\sum_i x_i y_i + 2a_1\sum_i x_i^2 + 2a_0\sum_i x_i = 0 \qquad (14)$$

que, cuando resolvemos para $a_0$ (ordenada al origen) y $a_1$ (pendiente) da:

$$a_0 = \frac{\left(\sum_i x_i^2\right)\sum_i y_i - \left(\sum_i x_i\right)\left(\sum_i x_i y_i\right)}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2} \qquad (15)$$

$$a_1 = \frac{n\sum_i x_i y_i - \left(\sum_i x_i\right)\left(\sum_i y_i\right)}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2} \qquad (16)$$

La ecuación para la línea de mejor ajuste se obtiene sustituyendo las ecuaciones 15 y 16 en la ecuación 7.

Nos hacemos ahora esta pregunta: "Cuáles son las incertezas en $a_0$ y $a_1$?" Cada $y_i$ tiene una incerteza (supuesta la misma para todos los $y_i$) y, entonces, ambas $a_0$ y $a_1$ tendrán incertezas. Estas incertezas son los desvíos estándar de las medias, $s_{ma0}$ y $s_{ma1}$. Para calcular $s_{ma0}$ y $s_{ma1}$, necesitamos el desvío estándar $s_y$.

Debemos responder entonces la pregunta previa: "Cual es la incerteza estadística en las medidas $y_1$, $y_2$,… $y_n$?". En este caso, el desvío estándar $s_y$ es

$$s_y = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - a_0 - a_1 x_i\right)^2} \qquad (17)$$

El desvío estándar de la media $s_{my}$ es

$$s_{my} = \frac{s_y}{n^{1/2}} \qquad (18)$$

Para cada $y_i$ el resultado a reportar es

$$y_i \pm s_{my} \qquad i = 1, 2, \ldots, n \tag{19}$$

La razón de la presencia del factor n-2 en el denominador de la ecuación 17 es que el cálculo de $a_0$ y $a_1$ reduce el número de pares de datos independientes ($x_i$, $y_i$) de n a n-2; el denominador en la ecuación para el desvío estándar es el número de pares de datos independientes.

*Nota*: Es importante verificar si los errores estimados, $dy_i$, registrados durante la toma de datos son consistentes con el error estadístico calculado $s_{my}$. Un desvío estándar de la media $s_{my}$, que sea mucho mayor que los errores estimados, $d_{yi}$, indicaría errores estimados que no fueron tenidos en cuenta. Errores experimentales, $d_{yi}$, que fueran mucho mayores que $s_{my}$ sugeriría un error estimado muy conservador, es decir, los valores $d_{yi}$ deberían haber sido estimados como valores más pequeños.

*EJERCICIO.*

Una investigadora en física planea calibrar su equipo determinando un valor medio para un parámetro x. Lo hace midiendo cuatro valores de x y estima el error δx. Suponga que los valores de x + δx son 2.741 ± 0.010, 2.832 ± 0.010, 2.678 ± 0.010, 2.763 ± 0.010. Calcule la media, $\overline{x}$ , y el desvío estándar de la media, $s_m$. Es su error estimado demasiado grande, demasiado pequeño o razonable? Explique su resultado.

Consideramos ahora los errores en $a_0$ y $a_1$, esto es: $s_{ma0}$, y $s_{ma1}$. Las ecuaciones 15 y 16 dan $a_0$ y $a_1$ como funciones de los valores medidos ($x_i$, $y_i$) donde el error estadístico para cada $y_i$ está dado en la ecuación 18. Como $a_0$ y $a_1$ son funciones conocidas de $y_i$ y los errores en $y_i$ son conocidos, los errores en $a_0$ y $a_1$ pueden ser determinados por propagación de errores. La fórmula básica para propagación de errores puede escribirse como

$$\delta Q = \sqrt{\sum_{j=1}^{n} \left( \frac{\partial Q}{\partial b_j} \right)^2 \left( \delta b_j \right)^2} \tag{20}$$

donde los valores medidos son $b_j \pm \delta b_j$, j=1,2,…,n, y dQ es el error en la cantidad calculada $Q(b_1, b_2, \ldots, b_n)$. Reemplazando δQ y $\delta b_j$ con desvíos estándar de la media $s_{mQ}$ y $s_{mbj}$ y elevando al cuadrado, tenemos

$$s_{mQ}^2 = \sum_{j=1}^{n} \left( \frac{\partial Q}{\partial b_j} \right)^2 s_{mbj}^2 \tag{21}$$

Aplicando la ecuación 21, $s_{ma0}$ es

$$s_{ma0}^2 = \sum_{j=1}^{n} \left( \frac{\partial a_0}{\partial y_i} \right)^2 s_{my}^2 \tag{22}$$

donde la derivada parcial $\partial a_0/\partial y_i$ se calcula usando la ecuación 15:

$$\frac{\partial a_0}{\partial y_j} = \frac{\sum_i x_i^2 - \left(\sum_i x_i\right) x_j}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2} \tag{23}$$

La ecuación 22 resulta, después de un poco de álgebra,

$$s_{ma0}^2 = \frac{s_{my}^2 \sum_i x_i^2}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2} \tag{24}$$

El resultado a ser reportado es

$$a_0 \pm s_{ma0} \tag{25}$$

El cálculo de $s^2_{ma1}$ es similar al cálculo de $s^2_{ma0}$. El resultado es

$$s_{ma1}^2 = \frac{n\, s_{my}^2}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2} \tag{26}$$

y reportamos

$$a_1 \pm s_{ma1} \tag{27}$$

**EJEMPLO.**

Estudio de la velocidad de un objeto (variable dependiente) como función del tiempo (variable independiente). Los datos son los siguientes:

| Velocidad (m/s) | tiempo(s) |
|---|---|
| 0.45 ± 0.06 | 1 |
| 0.81 ± 0.06 | 2 |
| 0.91 ± 0.06 | 3 |
| 1.01 ± 0.06 | 4 |
| 1.36 ± 0.06 | 5 |
| 1.56 ± 0.06 | 6 |
| 1.65 ± 0.06 | 7 |

| | |
|---|---|
| 1.85 ± 0.06 | 8 |
| 2.17 ± 0.06 | 9 |

Estos datos están graficados en la Figura 2.

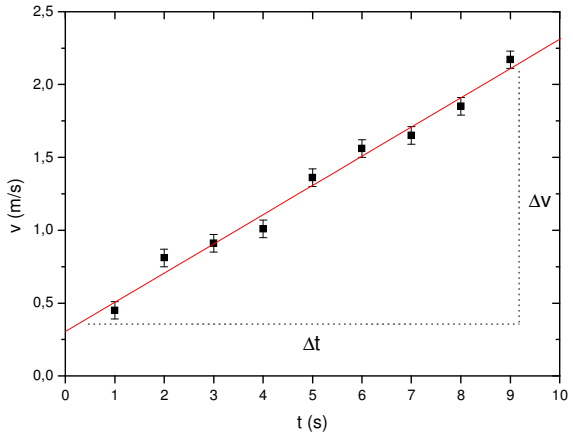


Figura 2

Velocidad versus tiempo. Los datos muestran una relación lineal.

Este gráfico muestra que la velocidad es una función lineal del tiempo. La ecuación general para una línea recta es

$$y = mx + b \qquad (28a)$$

Donde m es la pendiente de la línea y b, la intersección con el eje vertical, es el valor de y cuando x=0. Haciendo v = y, x = t, a = m, y $v_0$ = b,

$$v = at + v_0 \qquad (m/s) \qquad (28b)$$

Esta es la forma de la ecuación para la línea dibujada a través de los datos, donde $v_0$ es el valor de la velocidad a t = 0 y a es la pendiente de la línea, que es la aceleración del objeto. Del gráfico vemos que $v_0$ = 0.32 m/s. Para determinar la pendiente seleccionamos dos puntos sobre la línea (no sobre los datos) que estén bien separados, entonces

$$a = pendiente = \frac{\Delta v}{\Delta t} = \frac{2.35 - 0.40}{10.0 - 0.5}\frac{m/s}{s} = \frac{1.95}{9.5}\frac{m/s}{s} = 0.20\, m/s^2 \qquad (29)$$

La ecuación para la línea es

$$v = 0.20\, t + 0.32 \qquad m/s \qquad (30)$$

Aplicaremos el método de cuadrados mínimos y de regresión lineal a este problema.

La ordenada al origen $a_0$ (= $v_0$ , velocidad inicial) se calcula usando ecuación 15, donde n = 9, y el resultado es

$$a_0 = 0.305 \qquad m/s$$

La pendiente $a_1$ (= a, aceleración) se obtiene de la ecuación 16

$$a_1 = 0.201 \qquad m/s^2$$

Cuando $a_0$ y $a_1$ son conocidas, las ecuaciones 17 y 18 pueden ser usadas para calcular $s_{my}$:

$$s_{my} = 0.025 \qquad m/s$$

donde, en este caso, la variable dependiente y es la velocidad v. Para cada $v_i$, el resultado a reportarse es

$$v_i \pm s_{mv} = v_i \pm 0.025$$

Notar que $s_{mv}$ es menor que los errores estimados $dv_i$ = 0.06 m/s (ver tabla de datos), lo que sugiere que los errores estimados son muy conservadores o muy grandes.

Cuando $s_{mv}$ es conocido, las incertezas en $a_0$ ($s_{ma0}$) y $a_1$ ($s_{ma1}$) pueden ser calculadas usando las ecuaciones 24 y 26. Los resultados son

$$s_{ma0} = 0.018 \quad m/s$$
$$s_{ma1} = 0.003 \quad m/s^2$$

Entonces

$$a_0 \pm s_{ma0} = 0.305 \pm 0.018 \quad m/s \quad (velocidad\ inicial)$$
$$a_1 \pm s_{ma1} = 0.201 \pm 0.003 \quad m/s^2 \quad (aceleracion)$$

**Bibliografía**

(1) "The Art of Experimental Physics". D.W. Preston & Eric R. Dietz. John Wiley & Sons (1991). Pags. 18-26. Traducido y adaptado por José Luis Alessandrini (2014).