Regresión lineal por cuadrados mínimos simples y ponderados

Marcelo Otero

Regresión lineal por cuadrados mínimos ordinarios

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: $(x_i \pm \Delta x_i , y_i \pm \Delta y_i)$, con i de 1 hasta n. Supongamos además que hay motivos suficientes como para conjeturar que existe una relación funcional lineal entre X e Y (Figura 1, panel izquierdo), tal que:

$$Y = \mathbf{a} + \mathbf{b}X$$

donde **a** corresponde a la ordenada al origen y **b** a la pendiente de la recta.

En este caso, cabe preguntarse: ¿Me permiten mis datos afirmar que la relación lineal es correcta, al menos dentro del rango de valores medidos? En caso afirmativo, ¿cómo determino los parámetros **a** y **b**, y cuáles son sus incertezas?

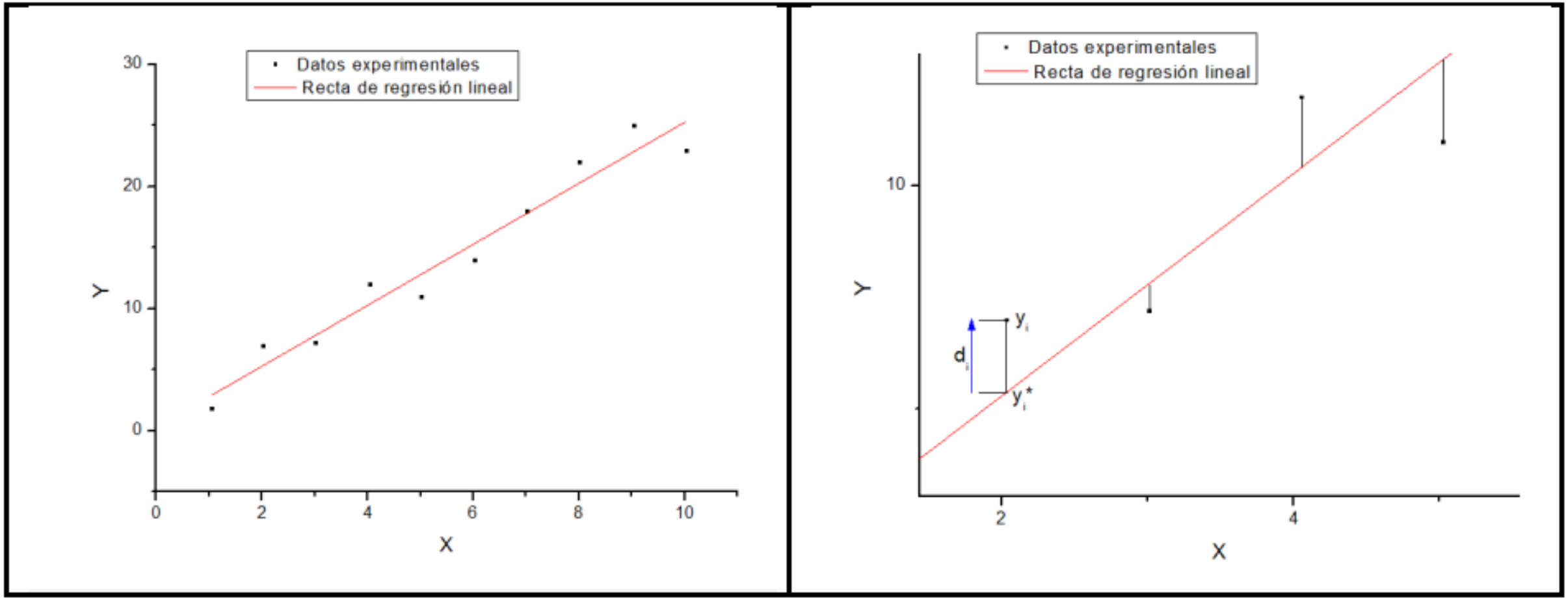


Figura 1: Panel izquierdo: datos experimentales y recta de regresión lineal. Panel derecho: visualización de $d_i = y_i - y_i^*$.

La técnica más usual para responder esas preguntas se conoce como el método de regresión lineal por cuadrados mínimos ordinarios, el cual supone despreciables las incertezas Δx_i y Δy_i.

¿Cómo determinar a y b? Para cada par de datos medido u observado (x_i,y_i), se define la cantidad: $d_i = y_i - (a+bx_i)$, donde d_i es la distancia entre el valor experimental observado y_i y el valor predicho por la recta y_i^* para el valor experimental xi, es decir $y_i^*=a+bx_i$ (Figura 1, panel derecho).

Es intuitivo ver que, si la relación es efectivamente lineal y los **a** y **b** son los correctos, cada d_i debe ser "pequeño". Como criterio para satisfacer esto, se puede proponer que la suma de todos los d_i sea mínima. Sin embargo, la suma de los d_i no es buen criterio, porque los d_i pueden ser positivos y negativos y cancelarse mutuamente. Para evitar este problema, se pueden sumar los módulos o, lo que es más conveniente, los d_i elevados al cuadrado: d_i^2. En el método de regresión lineal por **cuadrados mínimos ordinarios** se eligen los valores de **a** y **b** tales que minimicen la suma de los cuadrados de los d_i:

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

Es decir, la técnica de cuadrados mínimos ordinarios nos da expresiones analíticas para calcular la ordenada al origen **a** y la pendiente **b** de la recta de regresión lineal y una expresión para sus incertezas Δa y Δb respectivamente.

Queda todavía por contestar la pregunta de si los datos permiten asumir que las variables X e Y guardan una relación lineal. La técnica de regresión lineal por cuadrados mínimos ordinarios nos da un criterio para evaluar la confiabiliad o calidad de la relación lineal y es a través de un coeficiente, **r**, llamado coeficiente de correlación lineal de los datos.

El valor del índice de correlación lineal **r** varía en el intervalo [-1,1], indicando el signo el sentido de la relación:

- Si **r = 1**: **correlación lineal perfecta positiva** y los valores predichos coinciden con los observados, ya que todos los puntos de la nube están en la recta. Es decir, existe dependencia funcional que viene reflejada por una recta creciente.
- Si **r = -1**, la **correlación lineal es perfecta negativa** y, aquí también, los valores predichos coinciden con los observados, pero la recta es decreciente. De nuevo es un caso de dependencia funcional.
- Si **r = 0**, la **correlación lineal es nula**. Es decir, no hay asociación lineal y por mucho que varíe X, la variable Y no se verá afectada (de forma lineal).
- Si **-1 < r < 0**, la **correlación lineal será negativa** y la recta será decreciente puesto que el signo r coincide con el de la pendiente. Si r es cercano a 0 diremos que la relación es débil, y cuanto más se acerque a -1 consideráremos que la relación es más fuerte.
- Si **0 < r < 1**, la **correlación lineal es positiva**. Esto indica que la recta es creciente y cuando los valores de una variable crecen lo de la otra también crecerán. Consideraremos también que cuanto más se acerque a 0 más débil es la relación entre las variables y si el valor es próximo a 1 la relación podrá considerarse fuerte.

Regresión lineal por Cuadrados mínimos ponderados.

Se tiene un conjunto de datos experimentales: $(x_i \pm \Delta x_i, y_i \pm \Delta y_i)$, con i de 1 hasta n, y tal como en el caso anterior supongamos además que hay motivos suficientes como para conjeturar que existe una relación funcional lineal entre X e Y, tal que:

$$Y = \mathbf{a} + \mathbf{b}X$$

donde **a** corresponde a la ordenada al origen y **b** a la pendiente de la recta.

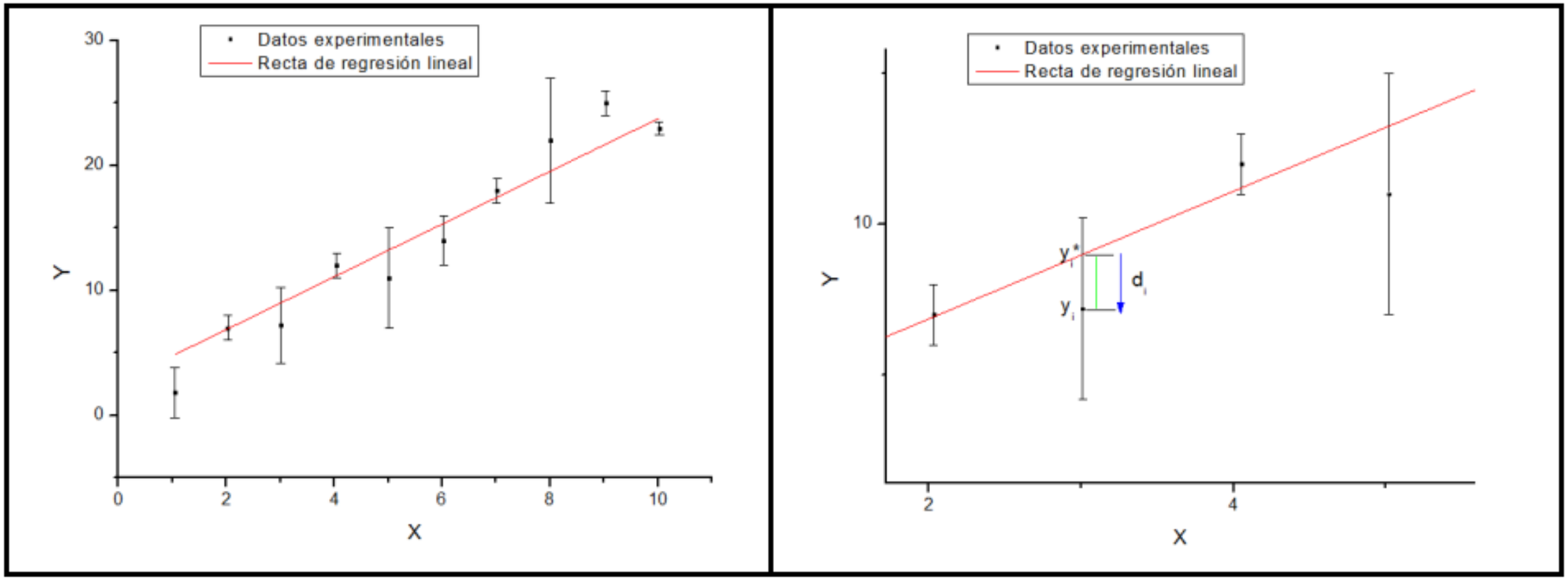


Figura 2: Panel izquierdo: datos experimentales con incerteza en Y, Δy_i, y recta de regresión lineal. Panel derecho: zoom en el que se visualiza d_i.

Antes de ver como se obtienen **a** y **b** usando un proceso de minimización equivalente al del caso anterior, vamos a dar un paso más para decidir que suma minimizar. Para ello, vamos a suponer que los errores de la magnitud X son despreciables frente a los errores de Y. Nótese, de paso, que d_i (Figura 2, panel derecho) es la distancia, en la dirección del eje y, entre y_i y el valor predicho por la recta $y_i^* = \mathbf{a} + \mathbf{b}x_i$. Es claro que, si los x_i "no tienen incerteza", la relación es lineal y sus parámetros son **a** y **b**, entonces las diferencias entre y_i e y_i^* son atribuibles en parte a las incertezas Δy_i. Pero si estas incertezas son distintas para cada y_i (Figura 2, panel izquierdo), parece sensato realizar el procedimiento de minimización asignando una mayor importancia a los d_i provenientes de valores de los y_i que tengan errores más chicos.

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 Δy_i. La forma más sencilla de hacer esto es dividir a cada d_i por Δy_i, tal que cuanto menor es la incerteza de y_i, mayor es el coeficiente que tiene el d_i. Usando este criterio, lo que se debe minimizar es:

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

Al método de obtención de **a** y **b** usando la minimización de Σ se lo llama "**cuadrados mínimos ponderados**".

Al igual que la técnica de cuadrados mínimos ordinarios, la técnica de cuadrados mínimos ponderados nos da expresiones analíticas para calcular la pendiente y la ordenada de la recta de regresión lineal y expresiones analíticas para sus incertezas: **a** +/- Δa y **b** +/- Δb. La técnica de regresión lineal por cuadrados mínimos ponderados nos da también un criterio para evaluar la confiabilidad o calidad de la relación lineal y es a través del mismo coeficiente, **r**, llamado coeficiente de correlación lineal de los datos.

Hipótesis y supuestos

Algunas ideas…

Ejemplos

	Conjunto 1		Conjunto 2	Conjunto 3	Conjunto 4	
N ° Observación.	x	y	y	y	x	y
1	10	8,04	9,14	7,46	8	6,58
2	8	6,95	8,14	6,77	8	5,76
3	13	7,58	8,74	12,74	8	7,71
4	9	8,81	8,77	7,11	8	8,84
5	11	8,33	9,26	7,81	8	8,47
6	14	9,96	8,1	8,84	8	7,04
7	6	7,24	6,13	6,08	8	5,25
8	4	4,26	3,1	5,39	19	12,5
9	12	10,84	9,13	8,15	8	5,56
10	7	4,82	7,26	6,42	8	7,91
11	5	5,68	4,74	5,73	8	6,89

- Algunos de estos puntos se ilustran mediante cuatro conjuntos de datos ficticios, cada uno de los cuales consta de once pares (x, y), que se muestran en la tabla. Para los primeros tres conjuntos de datos, los valores de x son los mismos y se enumeran solo una vez.
- Cada uno de los cuatro conjuntos de datos produce la misma salida estándar de un programa de regresión típico, a saber:
- Promedio de x= 9
- Promedio de y= 7.5
- Coeficiente b= 3
- Coeficiente a= 0.5
- R^2= 0,667

Ejemplo: Conjunto 1

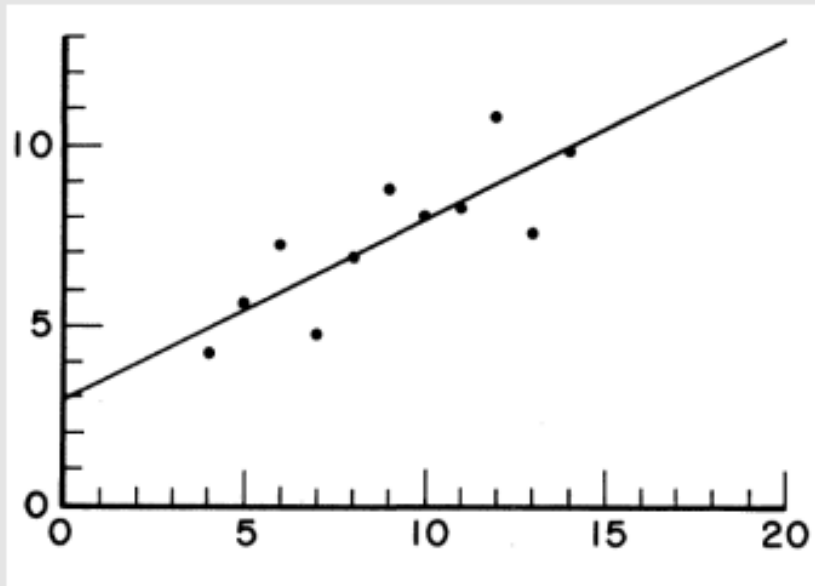


Figura 1. Conjunto de datos 1.

Ejemplo: Conjunto 2

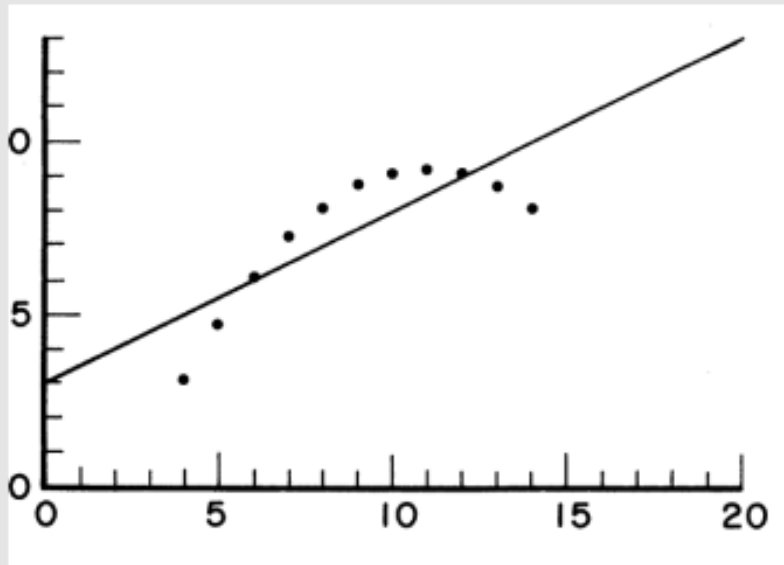


Figura 2. Conjunto de datos 2.

Ejemplo: Conjunto 3

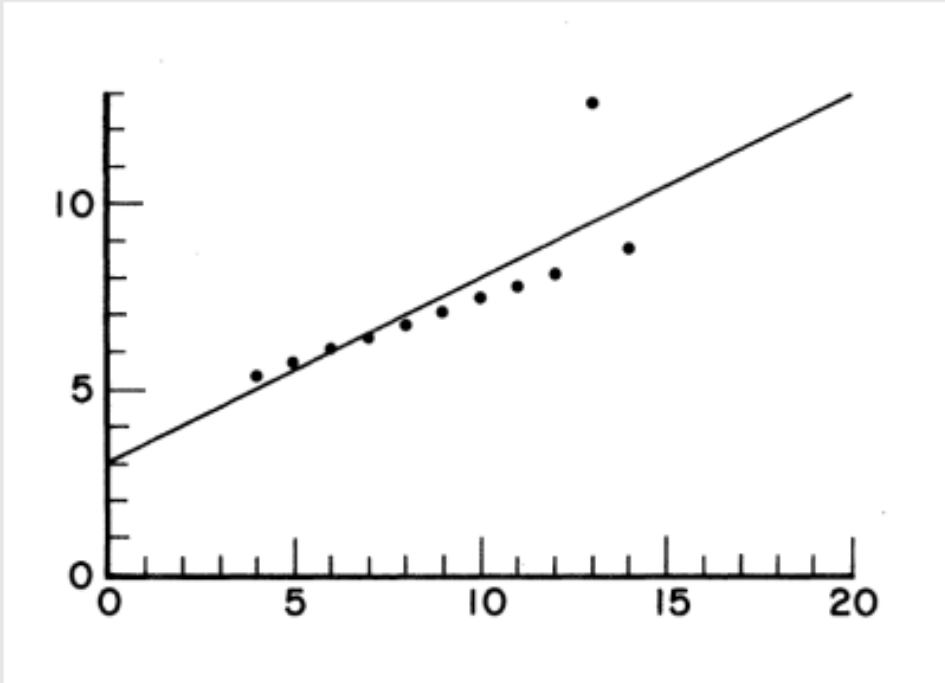


Figura 3. Conjunto de datos 3.

Ejemplo: Conjunto 4

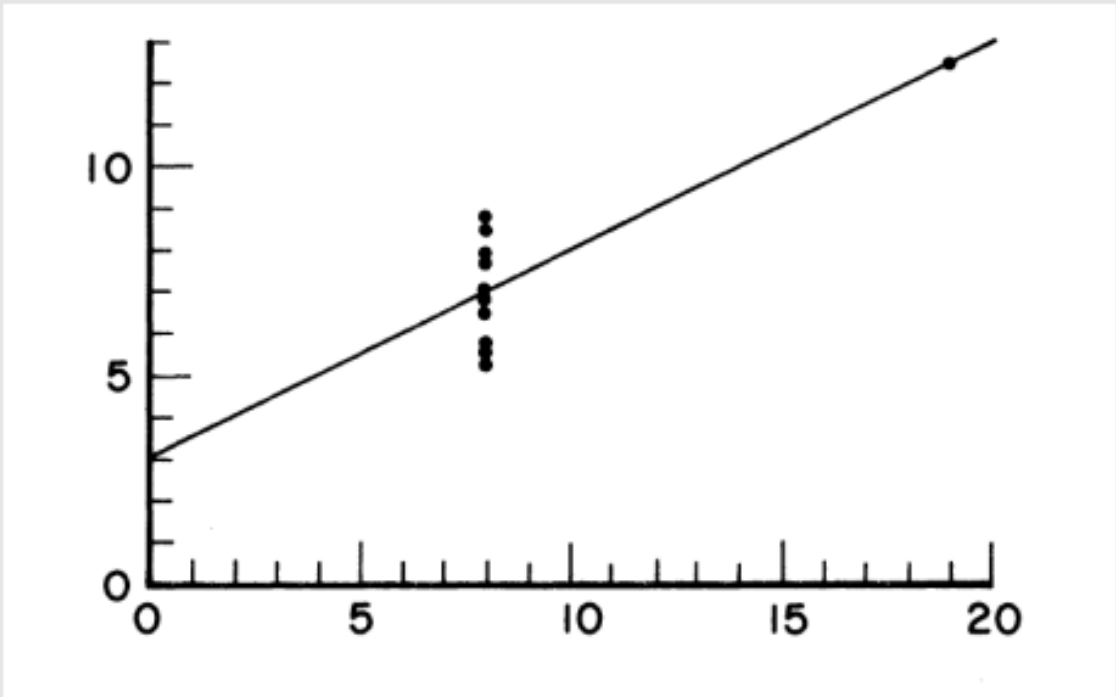


Figura 4. Conjunto de datos 4.

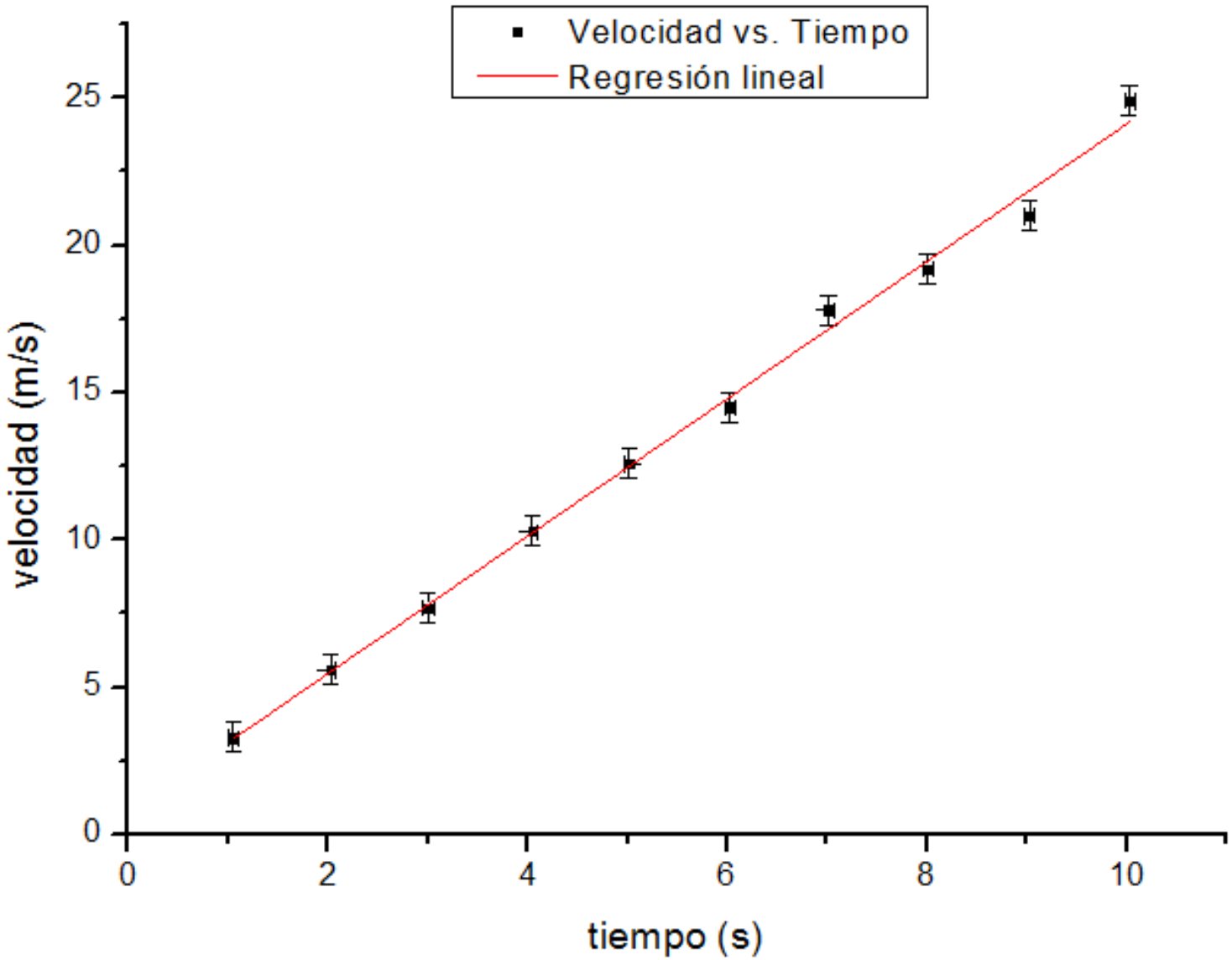
Ejemplos de cálculos de coeficientes a partir de los parámetros de Cuadrados Mínimos Ponderados (CMP)

Ejemplo 1

Se midió la velocidad de un móvil (V) a distintos tiempos (t) y se graficó la velocidad vs el tiempo
con sus respectivas incertezas. Teniendo en cuenta que se trata de un MRUV:

$$\mathbf{V = V_0 + \alpha * t}$$

Se quiere estimar la velocidad inicial (V_0) y la aceleración media del móvil (α).


Velocidad vs. Tiempo
Regresión lineal
velocidad (m/s)
tiempo (s)
0
5
10
15
20
25
0
2
4
6
8
10

Se realizó regresión lineal por cuadrados mínimos ponderados.

Modelo: y = **a** + **b**x

Coeficiente de correlación lineal de Pearson: **R=**0.99791
Coeficiente R^2=0.99529

Ambos son índices que miden la calidad del ajuste. El coeficiente de R de Pearson nos indica que tan buena es la correlación lineal entre las variables y toma valores en el intervalo [-1,1].

- Si R = 0 no hay correlación lineal.
- Si |R| =1 los datos son colineales.
- Si R>0 la pendiente es positiva.
- Si R<0 la pendiente es negativa.
- Cuanto más cercano a 1 es R, mayor es la correlación lineal entre las variables.

El R cuadrado mide la correlación lineal entre las variables pero toma valores en el intervalo [0,1]. Cuanto más cercano a 1 es el valor, mayor es la correlación lineal entre las variables.

Parámetros del ajuste (ordenada al origen y pendiente con sus incertezas):

	Valor	**Incerteza**
Ordenada al origen: a	0.80099	0.33275
Pendiente: b	2.33159	0.05343

En el informe, los nombres de los parámetros deben estar expresados con las cifras significativas adecuadas y con sus unidades.

Por ejemplo, considerando una sola cifra significativa en la incerteza:

La ordenada al origen es: a = (0.8 +/- 0.3) m/s
La pendiente es: b = (2.33 +/- 0.05) m/s2

En este ejemplo si las variables corresponden al tiempo y a la velocidad en un MRUV, entonces la regresión lineal (y= a + bx) se corresponde con la ecuación:

$$\mathbf{V = V_0 + \alpha * t}$$

donde V_0 es la velocidad inicial y α la aceleración. Por lo cual, en este ejemplo se podría estimar que

$$\mathbf{V_0 = a = (0.8 +/- 0.3)\ m/s}$$
$$\mathbf{\alpha = b = (2.33 +/- 0.05)\ m/s2}$$

Si nuestro objetivo era estimar la velocidad inicial V_0 y la aceleración α, en este caso los mismos coinciden con los parámetros de la regresión “a” y “b”.

Sin embargo, las magnitudes que uno quiere estimar no siempre coinciden con los parámetros de la regresión lineal “a” y “b”, sino que a veces son una función de los mismos (Ejemplo 2).

Ejemplo 2

Supongamos un experimento de termodinámica de gases ideales a temperatura constante (T) en el cual medimos la presión (P) de un gas para distintos volúmenes (V) siempre para un mismo número de moles (n) de un gas ideal. P, V, T y n son determinados experimentalmente y tienen asociados una incerteza (ΔP, ΔV, ΔT y Δn).

Nuestro objetivo es calcular la constante de los gases ideales R, para ello, hacemos 10 mediciones en el laboratorio de presión P para distintos volúmenes V y tenemos entonces 10 pares de datos (P+/-ΔP, V+/-ΔV).

Conociendo la ecuación de estado de un gas ideal, sabemos que la relación entre la presión y el volumen es **no-lineal (hiperbólica)** y está dada por la ecuación:

$$\mathbf{P} = \frac{nRT}{\boldsymbol{V}},$$ pero podemos reescribirla como: $\mathbf{P} = \mathrm{nRT}(\mathbf{1/V})$

Ésta no es ni más ni menos que una **relación lineal** entre la presión **P** y la inversa del volumen **(1/V)**, donde la pendiente corresponde a **nRT**.

Si realizamos una regresión lineal de los valores (P, 1/V) incluyendo sus incertezas: ΔP y $\Delta(1/V)$, esperamos obtener una recta de regresión:

$$y = a + bx$$

donde **y** corresponde a la variable **P** y **x** corresponde a la variable **1/V**.

Aclaración: $\Delta(1/V)$ se determina por propagación a partir de la incerteza $\Delta(V)$.

$$\Delta(1/V) = (1/V^2)\,\Delta(V)$$

La regresión lineal nos arrojará una ordenada a +/- Δa y una pendiente b +/- Δb, donde Δa y Δb corresponden a las incertezas de la ordenada al origen y la pendiente respectivamente.

Importante: "Siempre colocar como variable Y aquella con mayor error relativo"

En este caso:

$$y = a + bx \quad \text{equivale a} \quad P = nRT(1/V)$$

- La ordenada al origen debe valer cero, o lo que es equivalente, que **el valor cero debe estar contenido en el intervalo: $[a - \Delta a, a + \Delta a]$**

- La pendiente de la recta de regresión es $b=nRT$.

- Podemos despejar R de la expresión obteniendo: $R=b/nT$

Como vemos, la constante de los gases a determinar R, no coincide con la pendiente de la recta de regresión, sino que es función de la pendiente b y las variables n y T medidas experimentalmente.

- En este caso R = b/nT se calculará a partir de los valores de b, n y T observados y la incerteza de R (ΔR) se calculará por propagación de los errores de ΔT, Δn y Δb.

- R es una función de b, n y T que son tres magnitudes con incertezas, es decir:

R=f(b,n,t)

Utilizando la ecuación de propagación de errores (TP2):

Supongamos que se puede obtener en forma indirecta la magnitud W midiendo en forma directa las magnitudes x, y, z,... (independientes entre sí), mediante una función $f(x, y, z, ...)$, tal que $W = f(x, y, z, ...)$.

A partir de las mediciones directas, conocemos los valores: $x = x_0 \pm \Delta x$; $y = y_0 \pm \Delta y$; $z = z_0 \pm \Delta z$;....

Entonces, se puede obtener en forma indirecta la magnitud $W = W_0 \pm \Delta W$ siendo:

$$W_0 = f(x_o, y_o, z_o, ...) \tag{1}$$

$$\Delta W = \sqrt{\left[\frac{\partial f}{\partial x}(x_0, y_0, z_0, \dots) \cdot \Delta x\right]^2 + \left[\frac{\partial f}{\partial y}(x_0, y_0, z_0, \dots) \cdot \Delta y\right]^2 + \left[\frac{\partial f}{\partial z}(x_0, y_0, z_0, \dots) \cdot \Delta z\right]^2} \tag{2}$$

R = f(b,n,t) es equivalente a W = f(x,y,z) de la ecuación general de propagación de errores. Por lo cual ΔR se calcula como ΔW

Siendo R = b/nT

Para calcular ΔR, tengo que calcular las derivadas parciales: $\frac{\partial R}{\partial b}$, $\frac{\partial R}{\partial n}$ y $\frac{\partial R}{\partial T}$

$$\frac{\partial R}{\partial b} = \frac{1}{nT} \qquad \frac{\partial R}{\partial n} = \frac{-b}{Tn^2} \qquad \frac{\partial R}{\partial T} = \frac{-b}{nT^2}$$

Usando la ecuación de propagación de errores, entonces:

$$\Delta R = \sqrt{\left(\frac{\partial R}{\partial b}\Delta b\right)^2 + \left(\frac{\partial R}{\partial n}\Delta n\right)^2 + \left(\frac{\partial R}{\partial T}\Delta T\right)^2}$$

Finalmente se reporta R+/-ΔR con sus unidades correspondientes y con las cifras significativas adecuadamente expresadas, considerando 1 cifra significativa en la incerteza.

Ejemplo 3
Cálculo de g con un péndulo simple (TP3)

Mido el periodo (T) de un péndulo simple para distintos valores de longitud del hilo L

Objetivo: calcular g con su incerteza.

$$T = 2\pi\sqrt{\frac{L}{g}}$$

$$y = a + bx$$

Función lineal

Forma no lineal

T varía como la raíz cuadrada de L

$$T = 2\pi\sqrt{\frac{L}{g}}$$

$$\boldsymbol{T} = 2\pi\frac{\sqrt{\boldsymbol{L}}}{\sqrt{g}}$$

$$\boldsymbol{T^2} = 4\pi^2\frac{\boldsymbol{L}}{g}$$

Formas linealizadas

$$y = a + bx$$

$$T = 2\pi\sqrt{\frac{L}{g}}$$

$$T = 2\pi\frac{\sqrt{L}}{\sqrt{g}}$$

$$T = \frac{2\pi}{\sqrt{g}}\sqrt{L}$$

$$\sqrt{L} = \frac{\sqrt{g}}{2\pi}T$$

$$T^2 = 4\pi^2\frac{L}{g}$$

$$T^2 = \frac{4\pi^2}{g}L$$

$$L = \frac{g}{4\pi^2}T^2$$

Para poder hacer cuadrados mínimos ponderados se debe colocar en el eje Y aquella variable con mayor error relativo. De las cuatro linealizaciones se realizarán solo 2 regresiones lineales.

$$y = a + bx$$

$$T = 2\pi\sqrt{\frac{L}{g}} \Rightarrow T = 2\pi\frac{\sqrt{L}}{\sqrt{g}} \Rightarrow T = \frac{2\pi}{\sqrt{g}}\sqrt{L} \Rightarrow b_1 = \frac{2\pi}{\sqrt{g}} \Rightarrow g = \frac{4\pi^2}{{b_1}^2}$$

$$\sqrt{L} = \frac{\sqrt{g}}{2\pi}T \Rightarrow b_2 = \frac{\sqrt{g}}{2\pi} \Rightarrow g = 4\pi^2 {b_2}^2$$

$$T^2 = 4\pi^2\frac{L}{g} \Rightarrow T^2 = \frac{4\pi^2}{g}L \Rightarrow b_3 = \frac{4\pi^2}{g} \Rightarrow g = \frac{4\pi^2}{b_3}$$

$$L = \frac{g}{4\pi^2}T^2 \Rightarrow b_4 = \frac{g}{4\pi^2} \Rightarrow g = 4\pi^2 b_4$$

En todos los casos g = f(b), por lo cual la incerteza de g se obtiene por propagación de la expresión de g = f(b), no propagar g de la ecuación g = f(T, L)

$$T = 2\pi\sqrt{\frac{L}{g}}$$

$$T^2 = 4\pi^2 \frac{L}{g}$$

$$g = 4\pi^2 \frac{L}{T^2}$$

$$g = 4\pi^2 L T^{-2}$$

- g es lo que quiero calcular con su incerteza
- L es una magnitud medida en forma directa y la cual posee incerteza instrumental.
- T es una magnitud medida en forma directa como resultado de N mediciones y posee su incerteza absoluta (con contribución instrumental y estadística)
- π Es un número irracional con infinitos dígitos, el cual al truncarlo tiene asociada incerteza en el último dígito.

$$g = f(\pi, L, T)$$

$$\Delta g = \sqrt{\left(\frac{\partial g}{\partial \pi}\Delta\pi\right)^2 + \left(\frac{\partial g}{\partial L}\Delta L\right)^2 + \left(\frac{\partial g}{\partial T}\Delta T\right)^2}$$

$$\frac{\partial g}{\partial \pi} = 8\pi L T^{-2}$$

$$\frac{\partial g}{\partial L} = 4\pi^2 T^{-2}$$

$$\frac{\partial g}{\partial T} = 4\pi^2 L(-2)T^{-3} = -8\pi^2 L T^{-3}$$

Informamos:

$$g = g_0 \pm \Delta g$$

Conclusión

Para el caso del péndulo simple y considerando la ecuación 1:

Si tengo un único valor de periodo (proveniente de una única medición o un promedio de mediciones) para una ÚNICA longitud del hilo L, entonces despejo g de la ecuación 1, siendo g =f(T, L) y propago la incerteza de g a partir de las incertezas de T y L.

Si mido T (proveniente de una única medición o un promedio de mediciones) para distintas longitudes de hilo L, entonces busco una forma de linealización de la ecuación 1, obtengo g a partir de la pendiente "b" de esa forma lineal y obtengo la incerteza de g a partir de la propagación de incertezas de la pendiente b, teniendo en cuenta que g = f(b)

$$T = 2\pi\sqrt{\frac{L}{g}} \qquad (1)$$