Coeficientes de correlación y de determinación

Regresión lineal

Sirve para modelar la relación entre X y Y mediante una recta:

$$y = mx + b$$

Donde:

- **m** = pendiente
- **b** = intercepto
- **X** = variable independiente
- **Y** = valor estimado de la variable dependiente

La pendiente (m) indica cuánto cambia Y por cada unidad que cambia X.

El intercepto (b) es el valor de Y cuando X = 0 (el punto donde la recta corta el eje Y).

Método de los mínimos cuadrados

Busca la recta que minimiza la suma de los cuadrados de los errores verticales:

$$SSE = \sum \left(y_i - \hat{Y}_i\right)^2 = \sum e_i^2 \qquad e_i = y_i - \hat{Y}_i$$

Las fórmulas para los parámetros son:

$$m = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{\sum(x_i - \bar{X})^2} \qquad m = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$b = \bar{Y} - m\bar{X} \qquad b = \frac{\sum y - m\sum x}{n}$$

Ejemplo numérico

Datos:

X: 1, 2, 3

Y: 2, 3, 5

$$\sum x = 1 + 2 + 3 = 6$$

$$\sum y = 2 + 3 + 5 = 10$$

$$\sum xy = 2 + 6 + 15 = 23$$

$$\sum x^2 = 1 + 4 + 9 = 14$$

$$\sum y^2 = 4 + 9 + 25 = 38$$

$$n = 3$$

$$m = \frac{3(23) - 6(10)}{3(14) - (6)^2} = 1.5$$

$$b = \frac{10 - 1.5(6)}{3} = 0.33$$

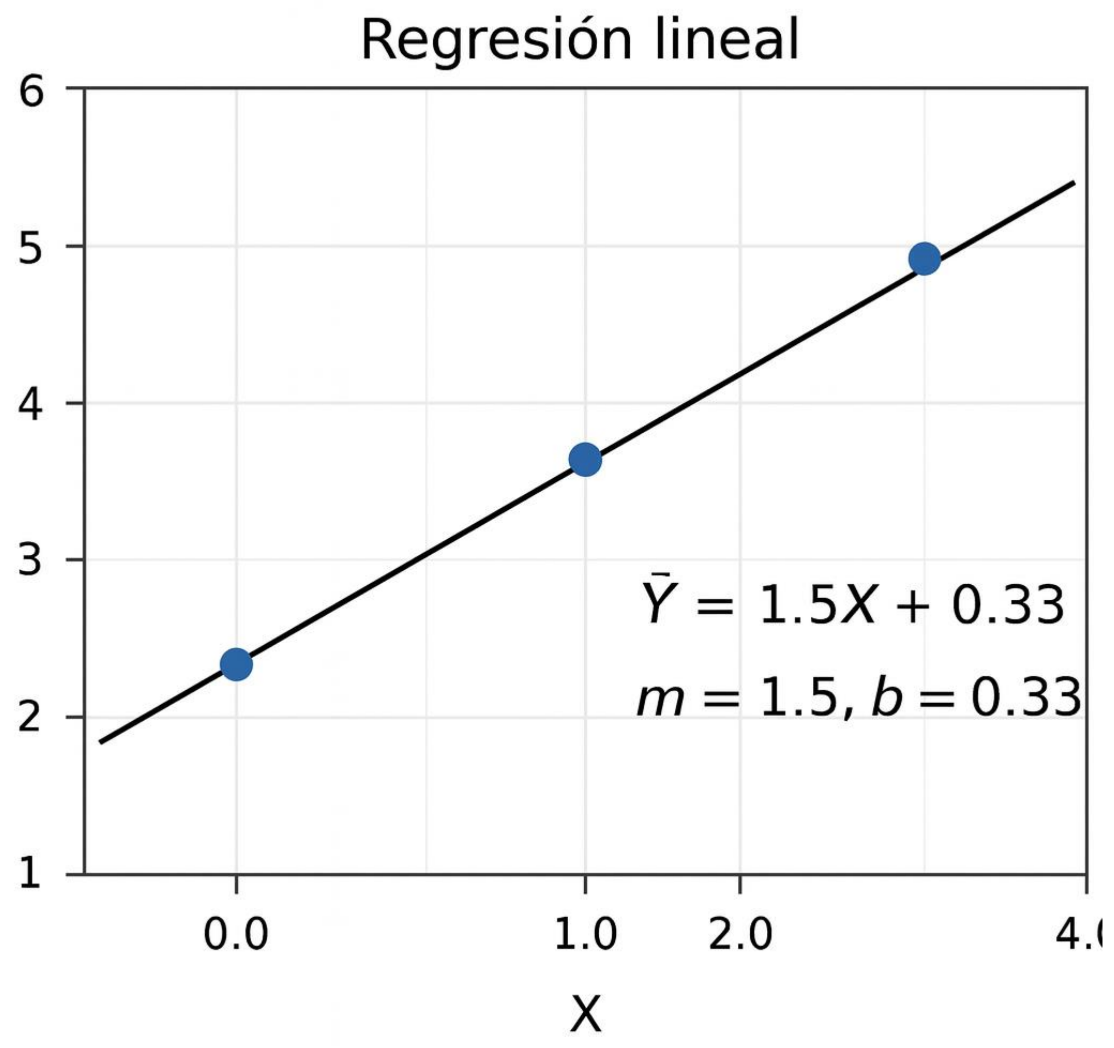
El método de mínimos cuadrados encuentra:

$$m = 1.5$$

$$b = 0.33$$

Ecuación de la recta final:

$$\hat{Y} = 1.5\,x - 0.33$$



Coeficiente de correlación *r*

El coeficiente clásico es: 👉 Coeficiente de correlación de Pearson (r)

Valor de r	Interpretación
0 a 0.3	Correlación débil
0.3 a 0.7	Correlación moderada
0.7 a 1.0	Correlación fuerte
0	No hay correlación
r negativo	Relación inversa

Fórmula general de r (Pearson):

$$r = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum(x_i - \bar{X})^2 \sum(y_i - \bar{Y})^2}}$$

Coeficiente de correlación *r*

$$r = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum(x_i - \bar{X})^2 \sum(y_i - \bar{Y})^2}}$$

El numerador mide **cómo varían juntas** las dos variables.

El denominador normaliza la expresión para que el resultado siempre quede entre –1 y 1.

Donde:

- x_i y y_i son los datos individuales
- $\bar{x}$ es la media de X
- $\bar{y}$ es la media de Y
- La suma se hace sobre todos los pares (x_i, y_i)

Coeficiente de correlación *r*

Fórmula práctica:

$$\boxed{r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}}$$

Esta fórmula es más fácil de usar con tablas de datos.

Coeficiente de determinación R^2

$$R^2 = r^2$$

Fórmula general en regresión lineal simple:

$$R^2 = 1 - \frac{SSE}{SST}$$

Donde:

SST es la suma total de cuadrados (variación total de Y)

SSE es la suma de cuadrados de los errores (lo que el modelo no explica)

$$SST = \sum (y_i - \bar{y})^2$$

$$SSE = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2$$

Ejemplo 1

Datos:

X = {2, 4, 6}

Y = {3, 6, 9}

1. Calcular el coeficiente de correlación.
2. Calcular el coeficiente de determinación.

Formula para calcular el coeficiente de correlación:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Formula para calcular el coeficiente de determinación:

$$R^2 = r^2$$

Ejemplo 1

Datos:

X = {2, 4, 6}

Y = {3, 6, 9}

$$\sum x = 2 + 4 + 6 = 12$$

$$\sum y = 3 + 6 + 9 = 18$$

$$\sum xy = 6 + 24 + 54 = 84$$

$$\sum x^2 = 4 + 16 + 36 = 56$$

$$\sum y^2 = 9 + 36 + 81 = 126$$

$$n = 3$$

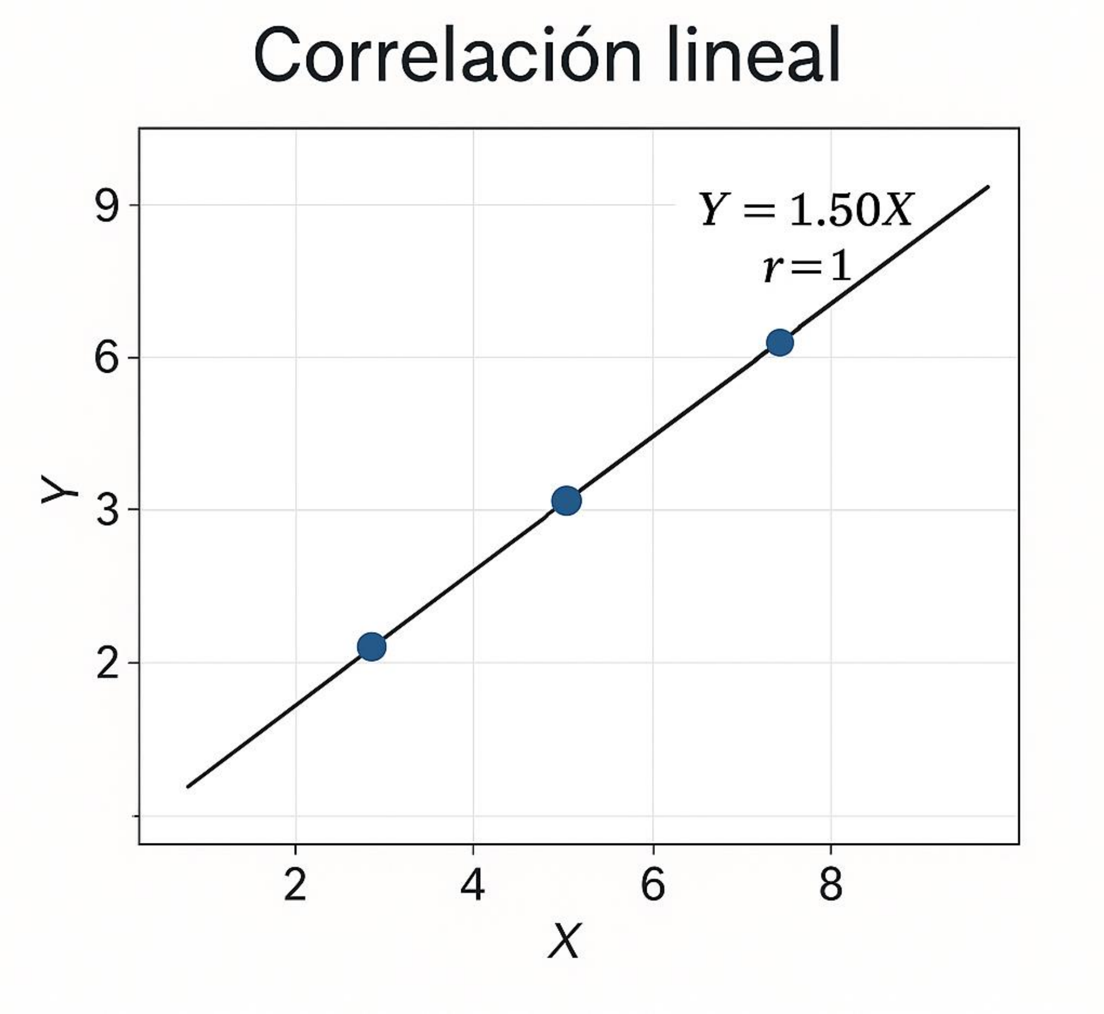
Ejemplo 1

	X	Y	XY	X^2	Y^2
	2	3	6	4	9
	4	6	24	16	36
	6	9	54	36	81
Sumas →	12	18	84	56	126

$$r = \frac{3(84) - (12)(18)}{\sqrt{[3(56) - (12)^2][3(126) - (18)^2]}} = \frac{36}{\sqrt{1296}} = \frac{36}{36} = 1$$

$$R^2 = r^2 = 1$$

Ejemplo 1



r = 1 → correlación lineal perfecta.

Ejemplo 2

En un laboratorio se midió la absorbancia de una solución de un colorante a distintas concentraciones:

Ensayo	Concentración (mg/L) X	Absorbancia (u.a.) Y
1	1	2.1
2	2	2.9
3	3	3.7
4	4	4.1
5	5	5.0

Se pide:

- Calcular la recta de regresión lineal de Y sobre X usando mínimos cuadrados.
- Interpretar los valores de m y b en el contexto.
- Estimar la absorbancia cuando la concentración es 3.5 mg/L.

Cálculo de la recta de regresión (mínimos cuadrados)

X	Y	XY	X^2
1	2.1	2.1	1
2	2.9	5.8	4
3	3.7	11.1	9
4	4.1	16.4	16
5	5.0	25.0	25

Sumas:

- $n = 5$
- $\sum X = 1 + 2 + 3 + 4 + 5 = 15$
- $\sum Y = 2.1 + 2.9 + 3.7 + 4.1 + 5.0 = 17.8$
- $\sum XY = 2.1 + 5.8 + 11.1 + 16.4 + 25.0 = 60.4$
- $\sum X^2 = 1 + 4 + 9 + 16 + 25 = 55$

Pendiente m: $$m = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$ $$m = \frac{35}{50} = 0.70$$

Intercepto b: $$b = \frac{\sum Y - m\sum X}{n}$$ $$b = \frac{7.3}{5} = 1.46$$

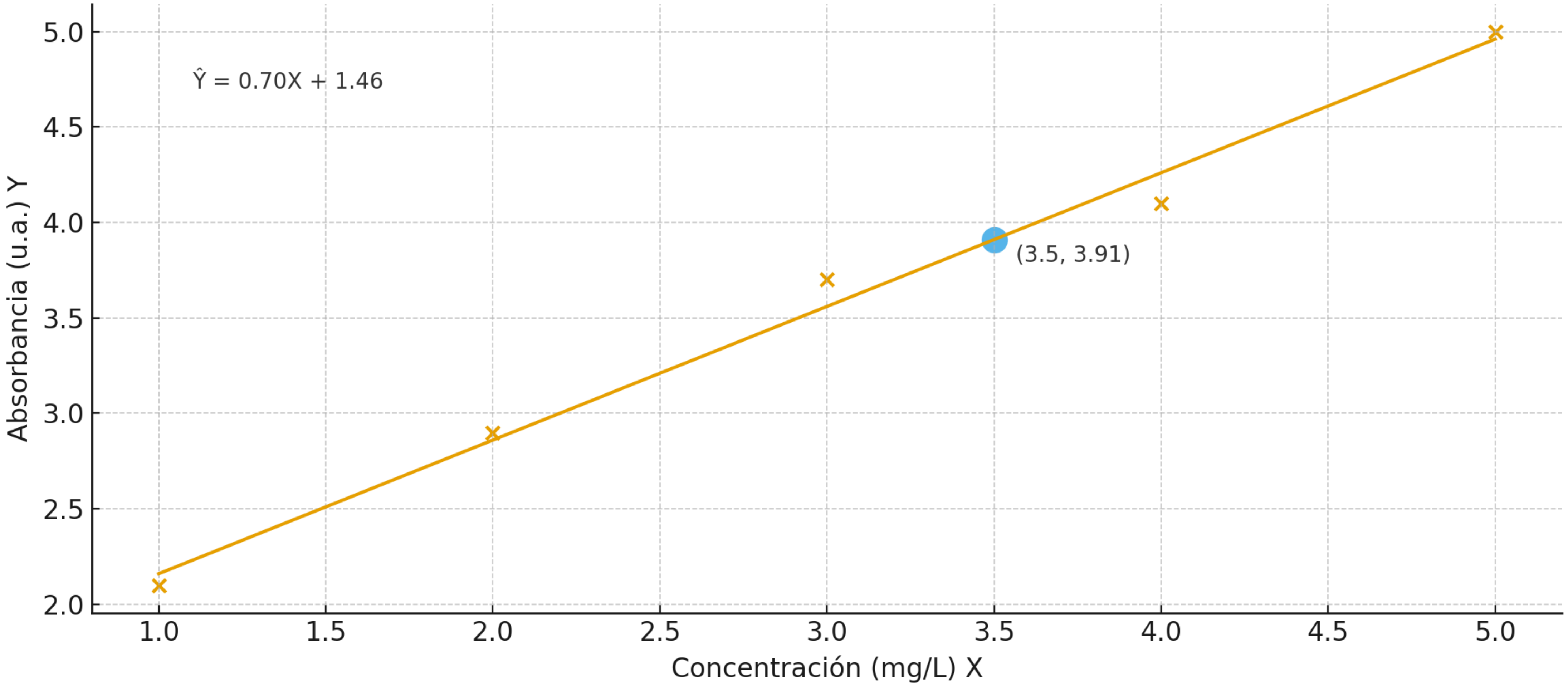
Ecuación final de la recta: $$\boxed{\hat{Y} = 0.70X + 1.46}$$

Predicción para X = 3.5 mg/L

Usamos la recta: $$\hat{Y} = 0.70X + 1.46$$

Para X = 3.5: $$\hat{Y} = 0.70(3.5) + 1.46$$ $$\boxed{\hat{Y}(3.5) \approx 3.91}$$

Respuesta: La absorbancia esperada a 3.5 mg/L es aproximadamente 3.91 unidades.



Ejemplo práctico. Curva de calibración de Fe^{2+}

Se obtiene la absorbancia de soluciones estándar de hierro en un espectrofotómetro UV-Vis.

Concentración (mg/L)	Absorbancia
0.20	0.09
0.40	0.20
0.60	0.31
0.80	0.40
1.00	0.51

Paso 1. Calculamos las medias:

$$\bar{X} = \frac{0.20 + 0.40 + 0.60 + 0.80 + 1.00}{5} = 0.60$$

$$\bar{Y} = \frac{0.09 + 0.20 + 0.31 + 0.40 + 0.51}{5} = 0.302$$

(X)	(Y)	XY	X^2
0.20	0.09	0.018	0.04
0.40	0.20	0.08	0.16
0.60	0.31	0.186	0.36
0.80	0.40	0.32	0.64
1.00	0.51	0.51	1.00
3.0	1.51	1.114	2.20

Paso 2. Calculamos la pendiente y el intercepto:

$$m = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{5(1.114) - 3.0(1.51)}{5(2.20) - (3.0)^2} = 0.52$$

$$b = \frac{\sum y - m\sum x}{n} = \frac{1.51 - (0.52)3.0}{5} = -0.01$$

Paso 3. Ecuación de la recta:

$$\hat{Y} = 0.52\,X - 0.01$$

Paso 5. Coeficiente de correlación (r)

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} = 0.998$$

El valor de r cercano a 1 indica una correlación lineal positiva casi perfecta.

Paso 6. Coeficiente de determinación

$$R^2 = r^2 = (0.998)^2 = 0.996$$

Interpretación: El 99.6% de la variabilidad de la absorbancia se explica por la concentración.

Ejemplo 1

(X)	1	2	3	4	5
(Y)	2.0	2.8	3.6	4.5	5.1

1. Ajustar la recta de regresión.
2. Calcular el coeficiente de correlación *r*.
3. Calcular el coeficiente de determinación R^2.
4. Interpretar el resultado con ayuda de las gráficas.

Calculamos las medias:

$$\bar{X} = \frac{1 + 2 + 3 + 4 + 5}{5} = 3$$

$$\bar{Y} = \frac{2.0 + 2.8 + 3.6 + 4.5 + 5.1}{5} = 3.6$$

Calculamos la pendiente:

$$m = \frac{\sum(X_i - \bar{X}\ )(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

Calculamos el intercepto:

$$b = \bar{Y} - m\bar{X}$$

Calculamos la pendiente:

$$m = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

(X)	(Y)	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
1	2.0	-2.0	-1.6	3.2	4.0
2	2.8	-1.0	-0.8	0.8	1.0
3	3.6	0.0	0.0	0.0	0.0
4	4.5	1.0	0.9	0.9	1.0
5	5.1	2.0	1.5	3.0	4.0
				7.9	10.0

$$m = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{7.9}{10.0} = 0.79$$

Calculamos el intercepto:

$$b = \bar{Y} - m\bar{X} = 3.6 - 0.79 \cdot 3 = 1.23$$

Ajustamos una recta $\hat{Y} = mx + b$ a esos datos.

Entonces la recta de regresión es: $\boxed{\hat{Y} = 0.79x + 1.23}$

Con ella calculamos los valores predichos $\widehat{Y}$ para cada X.

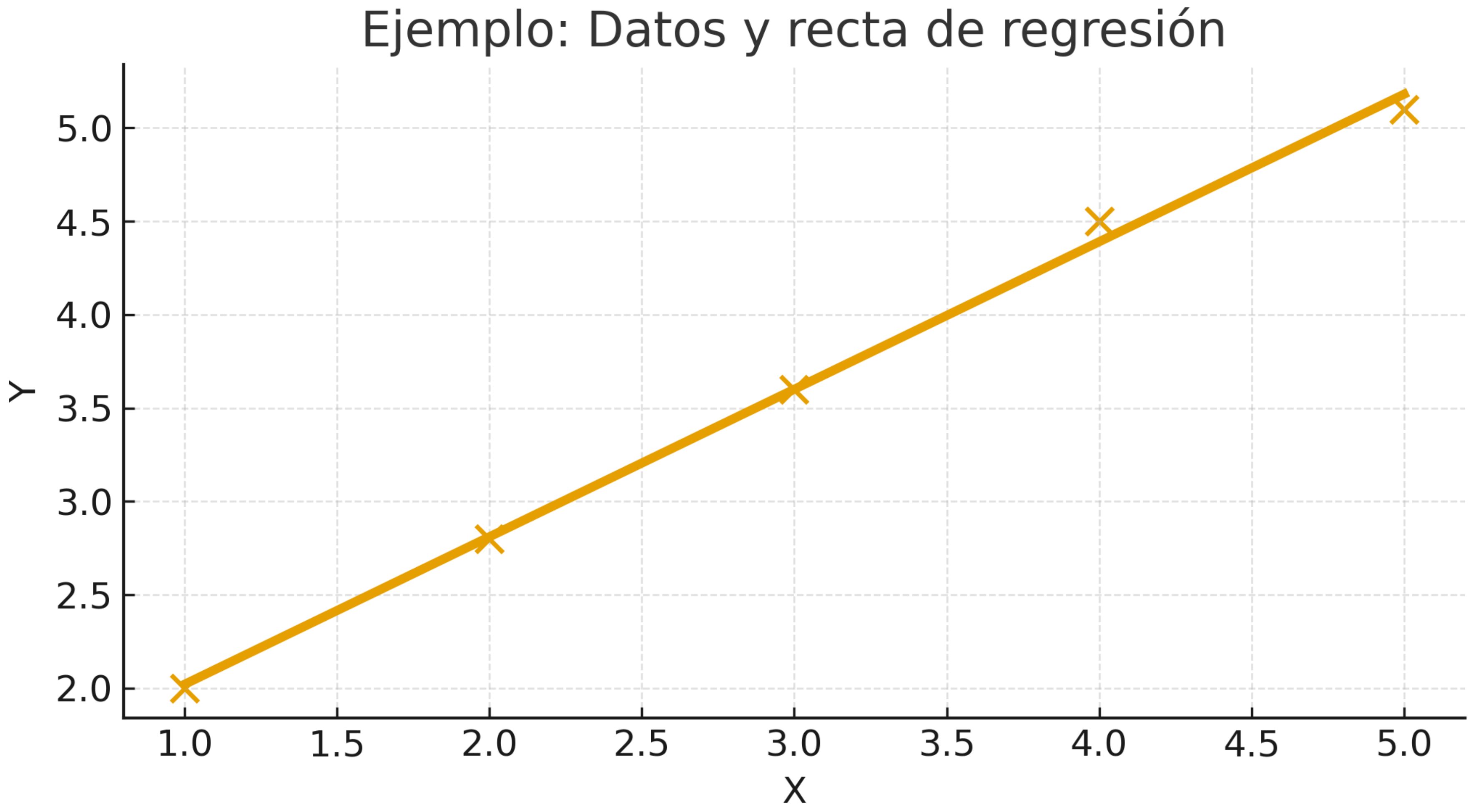
(X)	(Y)	$\hat{Y} = 0.79x + 1.23$
1	2.0	$\hat{Y} = 0.79\,(1) + 1.23 = 2.02$
2	2.8	$\hat{Y} = 0.79\,(2) + 1.23 = 2.81$
3	3.6	$\hat{Y} = 0.79\,(3) + 1.23 = 3.60$
4	4.5	$\hat{Y} = 0.79\,(4) + 1.23 = 4.39$
5	5.1	$\hat{Y} = 0.79\,(5) + 1.23 = 5.18$

Calculamos el coeficiente de correlación *r*:

$$r = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum(x_i - \bar{X})^2 \sum(y_i - \bar{Y})^2}}$$

(X)	(Y)	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	2.0	-2.0	-1.6	3.2	4.0	2.56
2	2.8	-1.0	-0.8	0.8	1.0	0.64
3	3.6	0.0	0.0	0.0	0.0	0.0
4	4.5	1.0	0.9	0.9	1.0	0.81
5	5.1	2.0	1.5	3.0	4.0	2.25
				7.90	10.0	6.26

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} = \frac{7.9}{\sqrt{10 \cdot 6.26}} = \frac{7.9}{7.912} = 0.9985$$


Ejemplo: Datos y recta de regresión
Y
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
X

Calculamos el coeficiente de determinación R^2.

$$R^2 = 1 - \frac{SSE}{SST}$$

Para eso, necesitamos los valores de SST y SSE:

$$SST = \sum (Y_i - \bar{Y})^2$$

$$SSE = \sum \left(Y_i - \hat{Y}_i\right)^2 = \sum e_i^2$$

(Y)	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$	Residuos: $e_i = Y_i - \hat{Y}_i$	e_i^2
2.0	-1.6	2.56	$2.00 - 2.02 = -0.02$	$(-0.02)^2 = 0.0004$
2.8	-0.8	0.64	$2.80 - 2.81 = -0.01$	$(-0.01)^2 = 0.0001$
3.6	0.0	0.00	$3.60 - 3.60 = 0.00$	$(0.00)^2 = 0$
4.5	0.9	0.81	$4.50 - 4.39 = 0.11$	$(0.11)^2 = 0.0121$
5.1	1.5	2.25	$5.10 - 5.18 = -0.08$	$(-0.08)^2 = 0.0064$
		6.26		0.019

$$SST = \sum (Y_i - \bar{Y})^2 = 6.26$$

$$SSE = \sum \left(Y_i - \hat{Y}_i\right)^2 = \sum e_i^2 = 0.019$$

Coeficiente de determinación:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{0.019}{6.26} = 0.997$$

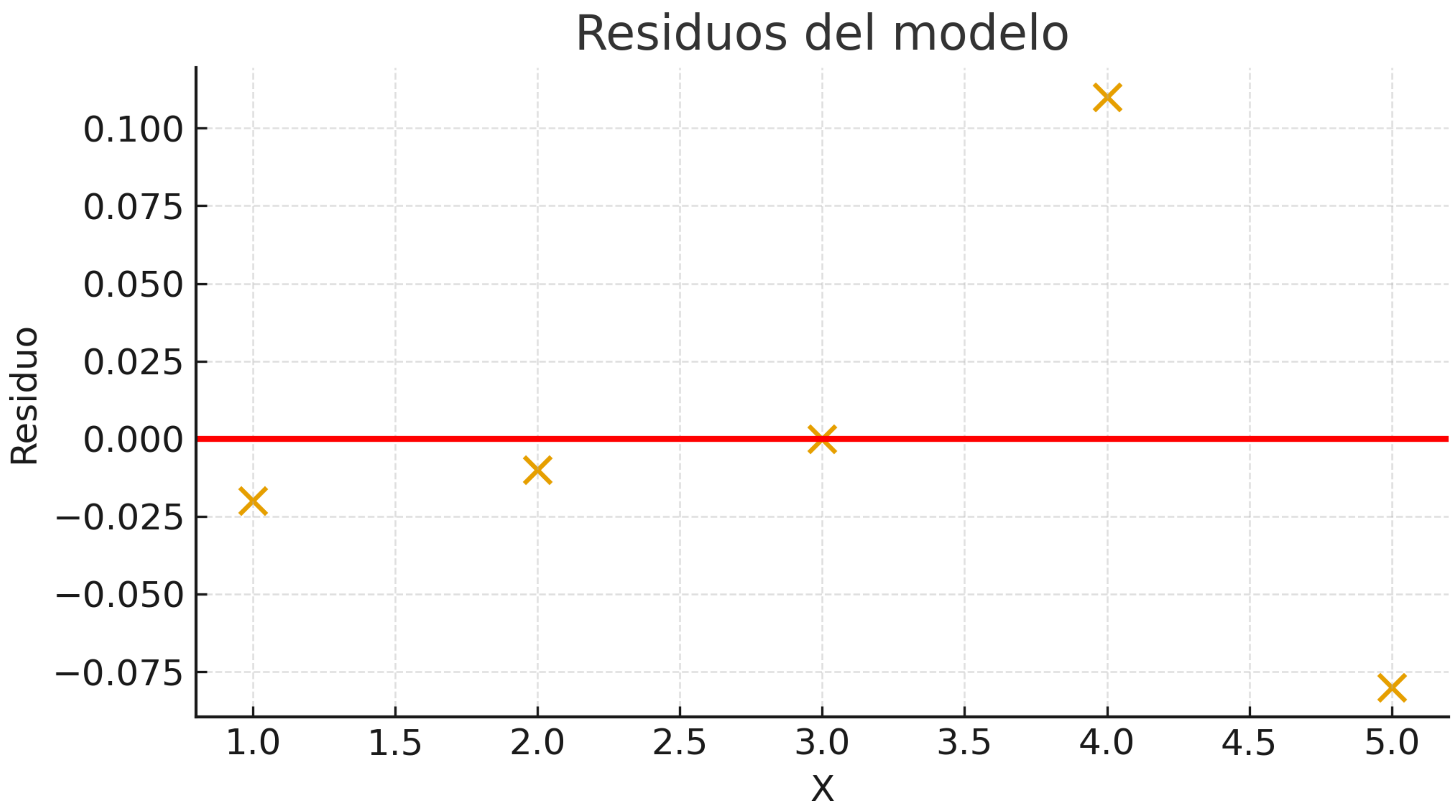
Interpretación

$R^2 = 0.997$ significa que:

- El 99.7 % de la variación de Y es explicada por la recta de regresión y solo el 0.3 % se debe a ruido o factores no incluidos en el modelo.
- Es un ajuste casi perfecto, típico de un ejemplo de libro o de una buena curva de calibración en laboratorio.

El coeficiente de determinación se puede también calcular directamente:

$$R^2 = r^2 = (0.9985)^2 = 0.997$$



Residuos del modelo
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
Residuo
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
X

Significación estadística del modelo lineal

Cuando ajustamos una regresión lineal:

$$\hat{Y} = mx + b$$

no basta con obtener la recta.

También debemos responder una pregunta fundamental:

¿La relación lineal entre X y Y existe realmente,

o es solo producto del azar?

A eso se le llama **significación estadística del modelo lineal**.

¿Qué significa que un modelo lineal sea “estadísticamente significativo”?

Significa que:

- La pendiente m es distinta de cero de manera suficientemente clara como para concluir que sí existe una relación entre X y Y en la población.

En términos matemáticos, se evalúa la hipótesis:

$$H_0 : m = 0 \quad \text{(no hay relación)}$$

$$H_a : m \neq 0 \quad \text{(sí hay relación)}$$

Si al realizar la prueba estadística obtenemos un valor **p pequeño**, típicamente:

$$p < 0.05$$

entonces **rechazamos** $\boldsymbol{H_0}$ y concluimos:

“La relación lineal entre X y Y es significativa.”

¿Cómo se evalúa esta significación?

Existen varios indicadores, y todos apuntan a la misma conclusión:

A) Prueba *t* para la pendiente

B) ANOVA de la regresión (prueba F)

C) Coeficiente de determinación R^2

A) Prueba t para la pendiente

La pendiente m calculada por mínimos cuadrados viene acompañada de:

- su error estándar $SE(m)$,
- un estadístico t,
- y su correspondiente valor p.

La fórmula general:

$$t = \frac{m}{SE(m)}$$

Un valor |t| grande → relación significativa.

$SE(m)$ es el **error estándar de la pendiente** del modelo lineal.

B) ANOVA de la regresión (prueba F)

Se utiliza la razón F:

$$F = \frac{Explicado\ por\ el\ modelo}{No\ explicado\ (error)}$$

Si $p_F < 0.05, \quad \rightarrow \ El\ modelo\ completo\ es\ significativo$

C) Coeficiente de determinación R²

Aunque R^2 no es una prueba en sí, sí indica cuánta variación explica el modelo.

Reglas aproximadas:

$$R^2 > 0.80 \rightarrow \ relación\ fuerte$$

$$R^2 \sim 0.50 \rightarrow \ relación\ moderada$$

$$R^2 < 0.20 \rightarrow relación\ débil, aunque\ podría\ ser\ significativa\ con\ n\ grande$$

⚠️ **OJO**:

Un R^2 bajo NO implica que la relación no sea significativa estadísticamente.

Ejemplo común: datos biológicos con mucha variabilidad natural.

Interpretación práctica

Si el modelo lineal es estadísticamente significativo, significa:

- Sí existe una relación real entre X y Y.
- La pendiente estimada (m) es confiable.
- El modelo puede hacerse generalizable a la población.
- Las predicciones basadas en la recta tienen fundamento estadístico.

Si NO es estadísticamente significativo:

- El modelo no es válido.
- X NO explica Y.
- La pendiente podría ser indistinguible de cero.
- No debemos usar la regresión para predecir.

Ejemplo sencillo (conceptual)

Supón que ajustas:

$$\hat{Y} = 3.2x + 2.1$$

y el software reporta:

- $m = 3.2$
- $SE(m) = 0.9$
- $t = 3.55$
- $p = 0.008$
- $R^2 = 0.78$

Interpretación:

- $p < 0.05 \rightarrow$ **la pendiente es significativa.**
- Existe una relación positiva entre X y Y.
- $R^2 = 0.78 \rightarrow$ 78% de la variación es explicada por el modelo.
- El modelo es útil para predicción.

Conclusión

La significación estadística del modelo lineal nos permite responder si:

- La relación entre X y Y es real, y no resultado del azar.
- Podemos confiar en la recta ajustada como una descripción válida del fenómeno.
- Es apropiado hacer predicciones usando ese modelo.
- Es el “sello de validez” de la regresión.

Ejemplo con datos reales (empresa de consumo)

Semana	Publicidad (X) (miles $)	Ventas (Y) (miles unidades)
1	5	55
2	8	60
3	12	63
4	15	70
5	18	74
6	22	80
7	25	82
8	30	90
9	35	95
10	40	102

La pregunta es: **¿La publicidad realmente explica las ventas?**

¿El modelo lineal es estadísticamente significativo?

Ajuste de la recta de regresión usando mínimos cuadrados

$$m = 1.25$$

$$b = 48.84$$

Ecuación de la recta final: $\hat{Y} = 1.25\,x + 48.84$

Interpretación: Cada 1 mil pesos adicionales en publicidad → las ventas aumentan 1,250 unidades.

Cálculo del coeficiente de determinación R^2

$$R^2 = 0.972$$

Interpretación: El 97.2% de la variación de las ventas es explicada por el gasto en publicidad. Esto ya sugiere una relación muy fuerte.

Significación estadística del modelo

El análisis estadístico del modelo lineal entrega:

- Estadístico F = 280.3
- p-value asociado = 0.0000008

Como: $p < 0.05$, Concluimos que:

El modelo lineal es estadísticamente significativo.

La relación entre publicidad y ventas sí existe en la población.

Gráfica de la relación (interpretación visual)

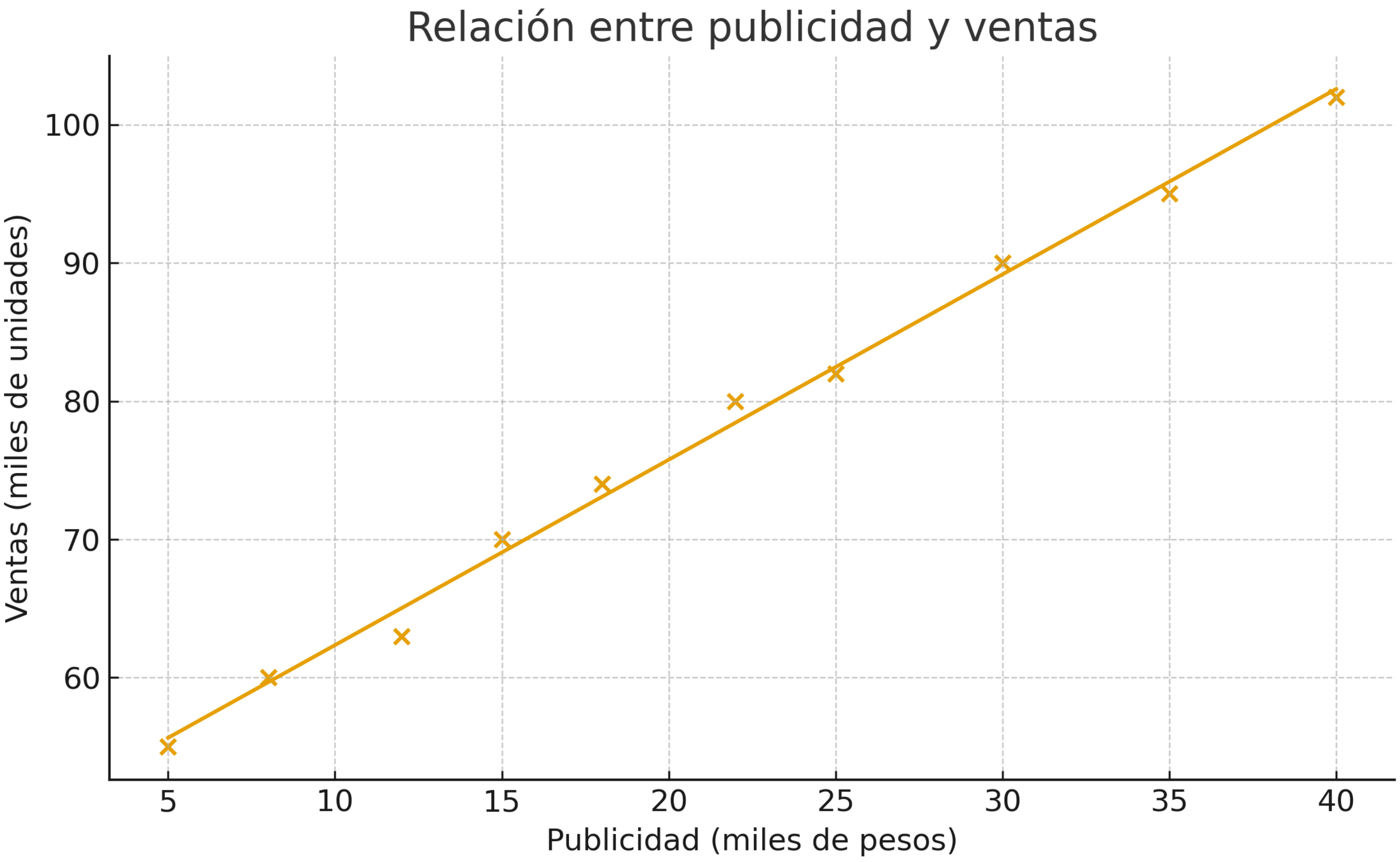


Tabla ANOVA del modelo (interpretada)

Fuente	SC (Suma de Cuadrados)	gl	CM	F	p-value
Modelo	2311.5	1	2311.5	280.3	0.0000008
Error	68.0	8	8.5	—	—
Total	2379.5	9	—	—	—

Interpretación:

- Casi toda la variación está en la parte explicada por el modelo.
- La F es enorme → evidencia muy fuerte de relación lineal.

Conclusión del ejemplo

- La publicidad sí explica las ventas.
- El modelo lineal no es producto del azar (p muy pequeño).
- $R^2 = 0.972$ indica una relación muy fuerte.
- El modelo puede usarse para predicción con alta confianza.

<u>Ejemplo de predicción:</u>

Si se invierten 50 mil pesos en publicidad:

$$\hat{Y} = 1.25\ (50) + 48.84 = 110.34$$

Se esperan aprox. 110 mil unidades vendidas.

Ejemplo: Curva de calibración espectrofotométrica

Se quiere cuantificar un fármaco por espectrofotometría UV–Vis a una longitud de onda específica. Se preparan soluciones estándar y se mide la absorbancia.

Datos experimentales:

Ensayo	Concentración (X) (mg/L)	Absorbancia (Y) (u.a.)
1	0	0.02
2	1	0.19
3	2	0.40
4	3	0.61
5	4	0.79
6	5	1.02
7	6	1.19

Ajuste del modelo lineal

Se ajusta una recta de calibración:

$$\hat{Y} = mx + b$$

Usando mínimos cuadrados (por ejemplo en Excel /otro software), se obtiene:

$$m = 0.199$$

$$b = 0.007$$

Ecuación de la recta final: $\hat{Y} = 0.199\,x + 0.007$

Interpretación de m: Por cada incremento de 1 mg/L en concentración, la absorbancia aumenta aproximadamente 0.199 unidades.

Coeficiente de determinación: $R^2 = 0.999$

Interpretación: El 99.9 % de la variación de la absorbancia se explica por la concentración. Es un ajuste excelente, típico de una buena curva de calibración en su rango lineal.

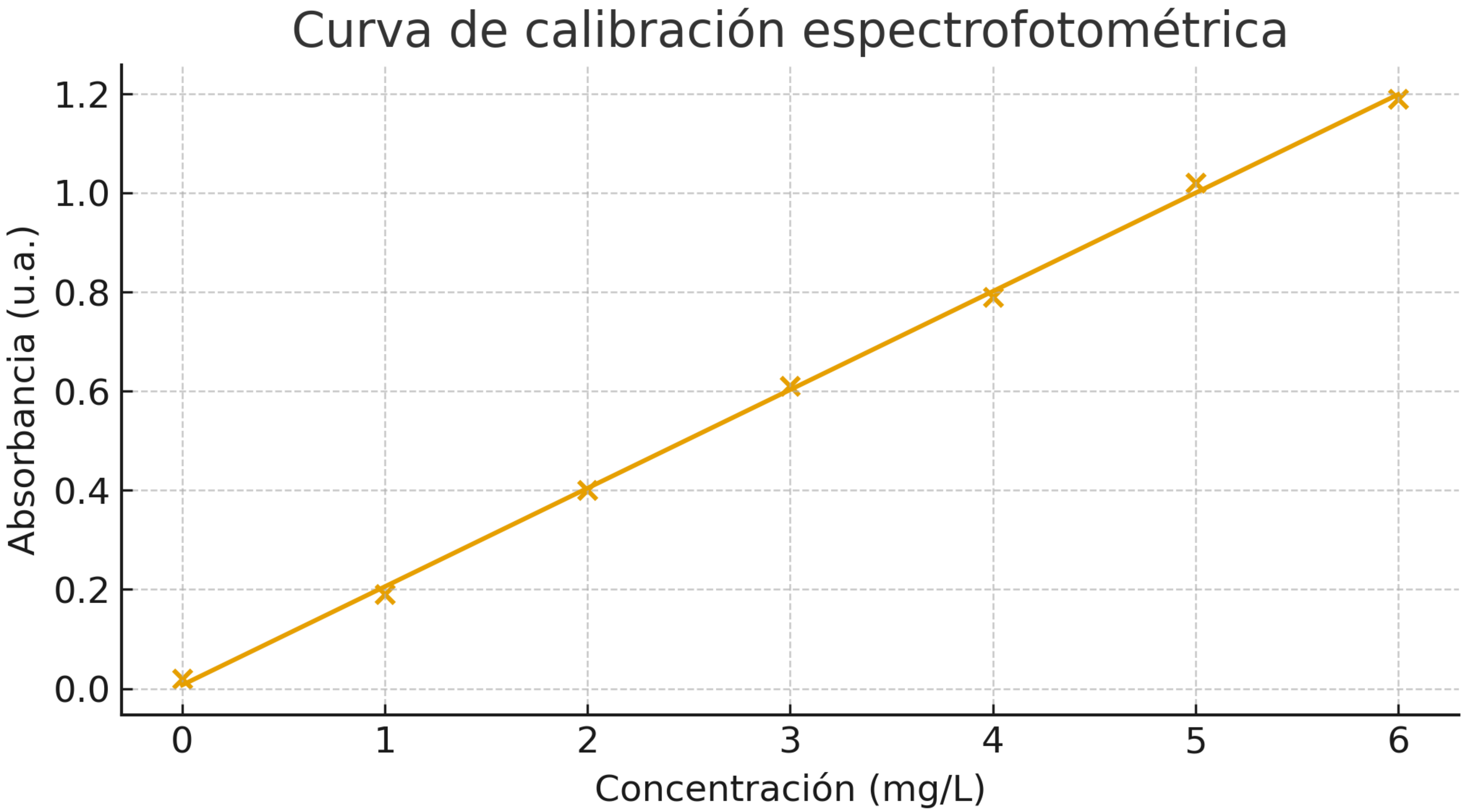
Significación estadística del modelo

En un software de análisis (Excel, Origin, etc.) verías:

- p-valor de la pendiente $m \ll 0.05$
- F de la regresión muy grande

Conclusión:

- El modelo lineal es estadísticamente significativo.
- La relación "concentración–absorbancia" es real, no producto del azar.
- Por lo tanto, se puede usar la ecuación para determinar concentraciones desconocidas a partir de su absorbancia.



Curva de calibración espectrofotométrica
Absorbancia (u.a.)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
1
2
3
4
5
6
Concentración (mg/L)

Curva de calibración obtenida

$$\hat{Y} = 0.199\,x + 0.007$$

Supongamos que se mide la absorbancia de una muestra desconocida y se obtiene:

$$A_{muestra} = 0.72$$

Necesitamos encontrar su concentración. $A = \pi r^2 —$

Sustituimos en la ecuación:

$$A = 0.199\,x + 0.007$$

$$0.72 = 0.199\,x + 0.007$$

$$x = 3.58\,mg/L$$

Resultado

$$X_{muestra} = 3.58\, mg/L$$

Verificación rápida

$$\hat{Y} = 0.199\,(3.58) + 0.007 = 0.713 + 0.007 = 0.720$$

Ejemplo con muestras fuera del rango lineal.

Ensayo	Concentración (X) (mg/L)	Absorbancia (Y) (u.a.)
1	0	0.02
2	1	0.19
3	2	0.40
4	3	0.61
5	4	0.79
6	5	1.02
7	6	1.19

$$\hat{A} = 0.199\,x + 0.007$$

El laboratorio valida que la región lineal útil de la curva es: $0 \leq X \leq 6\,mg/L$

Se mide la absorbancia de una muestra problema sin diluir y se obtiene:

$$A_{muestra} = 1.45$$

Pregunta:

- ¿Es válido usar directamente la ecuación de calibración para calcular la concentración a partir de $\mathrm{A} = 1.45$?

Si ignoráramos el rango lineal y usáramos:

$$\hat{A} = 0.199\,x + 0.007 \qquad \rightarrow \qquad X \sim 7.25\,mg/L$$

PEROOOO… El rango lineal validado es hasta 6 mg/L.

Estamos **fuera del rango lineal** de la curva.

Conclusión:

- No es válido confiar en este valor.
- El instrumento y el sistema ya no garantizan linealidad arriba de 6 mg/L.
- El resultado directo de 7.25 mg/L es **sospechoso y potencialmente sesgado.**

Prueba de hipótesis e intervalos de confianza para los parámetros de la línea recta.

Este es uno de los temas más importantes en regresión lineal, porque permite:

- Evaluar si la relación entre X y Y realmente existe.
- Determinar si la pendiente es distinta de cero.
- Estimar la recta verdadera de la población con un intervalo de confianza.

1. Modelo de regresión lineal simple

Trabajamos con el modelo:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Donde:

$$\beta_0 = intercepto\ poblacional$$
$$\beta_1 = pendiente\ poblacional$$
$$\varepsilon = error\ aleatorio$$

En la muestra obtenemos los estimadores:

$$b_0 \text{ para } \beta_0$$
$$b_1 \text{ para } \beta_1$$

2. Prueba de hipótesis para la pendiente β_1

La pregunta es: **¿Existe relación lineal entre X y Y en la población?**

Esto se evalúa mediante una **prueba t**:

$$H_0: \quad \beta_1 = 0 \quad \text{(no hay relación lineal)}$$
$$H_1: \quad \beta_1 \neq 0 \quad \text{(si hay relación lineal)}$$

Estadístico *t*: $$t = \frac{b_1}{SE(b_1)}$$

$SE(b_1)$ es el error estándar de la pendiente

Decisión

- Si $|t|$ es grande → evidencia contra H_0
- Si $p < \alpha$ (generalmente 0.05) → **se rechaza** H_0

Conclusión

Si se rechaza H_0:

> Existe una relación lineal significativa entre X y Y.

3. Prueba de hipótesis para el intercepto β_0

$$H_0: \quad \beta_0 = 0$$
$$H_1: \quad \beta_0 \neq 0$$

$$t = \frac{b_0}{SE(b_0)}$$

- Aunque es menos importante que la pendiente, se utiliza para validar ciertos modelos
- (p. ej., curvas de calibración que se “obligan” a pasar por el origen).

4. Intervalos de confianza para los parámetros

Los parámetros estimados b_0 y b_1 **NO son exactos**:

son estimaciones con variabilidad muestral.

Por eso se construyen intervalos de confianza:

Intervalo de confianza del 95% para la pendiente $\beta 1$

$$b_1 \pm t_{\alpha/2,\, n-2} \cdot SE(b_1)$$

Interpretación:

El 95% de todas las rectas poblacionales posibles tienen pendientes dentro de este intervalo.

Intervalo de confianza del 95% para el intercepto $\beta 0$

$$b_0 \pm t_{\alpha/2,\, n-2} \cdot SE(b_0)$$

Interpretación general de los intervalos

Para la pendiente:

- Si el intervalo incluye 0, entonces la pendiente podría ser 0 → NO hay evidencia de relación lineal.
- Si el intervalo NO incluye 0, entonces la pendiente es distinta de 0 → SÍ hay relación lineal significativa.

Para el intercepto:

- Si incluye 0, no es necesariamente un problema.
- En calibración analítica, un intercepto muy lejos de 0 indica problemas del método.

Ejemplo numérico

Supongamos que en un análisis se obtuvo:

- $b_1 = 2.5$
- $SE(b_1) = 0.6$
- $n = 12$ puntos experimentales
- Nivel de confianza 95%

Con $gl = n - 2 = 10$:

$$t_{0.025,10} = 2.228$$

Intervalo de confianza:

$$IC = 2.5 \pm 2.228(0.6)$$

$$IC = 2.5 \pm 1.3368$$

$$IC = (1.163, 3.837)$$

Interpretación:

- La pendiente **no incluye 0** →
  Existe una relación lineal significativa.
- La verdadera pendiente poblacional se encuentra entre **1.16 y 3.84** con 95% de confianza.

Ejemplo. Datos de calibración

Se prepara una serie de soluciones estándar de un analito y se mide su absorbancia UV–Vis:

Ensayo	Concentración (X) (mg/L)	Absorbancia (Y) (u.a.)
1	0.0	0.01
2	1.0	0.21
3	2.0	0.39
4	3.0	0.61
5	4.0	0.80
6	5.0	1.02

1. Ajustar el modelo lineal $\widehat{Y} = m\,x + b$.
2. Hacer prueba de hipótesis para la pendiente $\beta 1$.
3. Hacer prueba de hipótesis para el intercepto $\beta 0$.
4. Construir intervalos de confianza al 95 % para $\beta 1$ y $\beta 0$.
5. Ver la gráfica.

1. Ajuste de la recta de regresión

Con mínimos cuadrados se obtiene:

- Pendiente estimada:

$$m = b_1 \approx 0.2011$$

- Intercepto estimado:

$$b = b_0 \approx 0.0038$$

Ecuación de regresión:

$$\hat{Y} = 0.2011X + 0.0038$$

Interpretación de m:

Cada aumento de **1 mg/L** en concentración incrementa la absorbancia en aproximadamente **0.201 unidades**.

2 Calidad del ajuste: R^2

Calculando SST, SSR y SSE se obtiene:

$$R^2 \approx 0.9993$$

> El **99.93 %** de la variación de la absorbancia se explica por la concentración.
> Excelente linealidad, apropiada para una curva de calibración.

3 Prueba de hipótesis para la pendiente β_1

Hipótesis

$$H_0 : \beta_1 = 0 \quad \text{(no hay relación lineal)}$$

$$H_a : \beta_1 \neq 0 \quad \text{(sí hay relación lineal)}$$

Error estándar de la pendiente

$$SE(b_1) \approx 0.00270$$

Estadístico t

$$t_{\text{calc}} = \frac{b_1}{SE(b_1)} = \frac{0.2011}{0.00270} \approx 74.48$$

Con $gl = 4$ y $\alpha = 0.05$ (bilateral), el valor crítico es:

$$t_{0.025,4} \approx 2.776$$

Como:

$$|t_{\text{calc}}| = 74.48 \gg 2.776$$

⇒ **se rechaza H_0.**

📌 **Conclusión:**

La pendiente es altamente significativa.

Existe una relación lineal muy fuerte entre concentración y absorbancia.

4 Prueba de hipótesis para el intercepto β_0

Hipótesis

$$H_0 : \beta_0 = 0$$

$$H_a : \beta_0 \neq 0$$

Error estándar del intercepto

$$SE(b_0) \approx 0.00818$$

Estadístico t

$$t_{\text{calc}} = \frac{b_0}{SE(b_0)} = \frac{0.0038}{0.00818} \approx 0.466$$

Comparamos con el mismo valor crítico $t_{0.025,4} = 2.776$:

$$|t_{\text{calc}}| = 0.466 < 2.776$$

⇒ **NO se rechaza H_0.**

📌 **Conclusión**:

El intercepto **no es significativamente diferente de 0**.

Es razonable considerar que, a concentración cero, la absorbancia real podría ser muy cercana a cero (compatible con la ley de Beer-Lambert).

5 Intervalos de confianza al 95 %

a) Para la pendiente β_1

$$IC_{95\%}(\beta_1) = b_1 \pm t_{0.025,4} \cdot SE(b_1)$$

$$IC_{95\%}(\beta_1) = 0.2011 \pm 2.776 \times 0.00270$$

$$IC_{95\%}(\beta_1) \approx (0.1936,\ 0.2086)$$

✔ La pendiente es claramente positiva y bien acotada; el intervalo **no incluye 0**.

b) Para el intercepto β_0

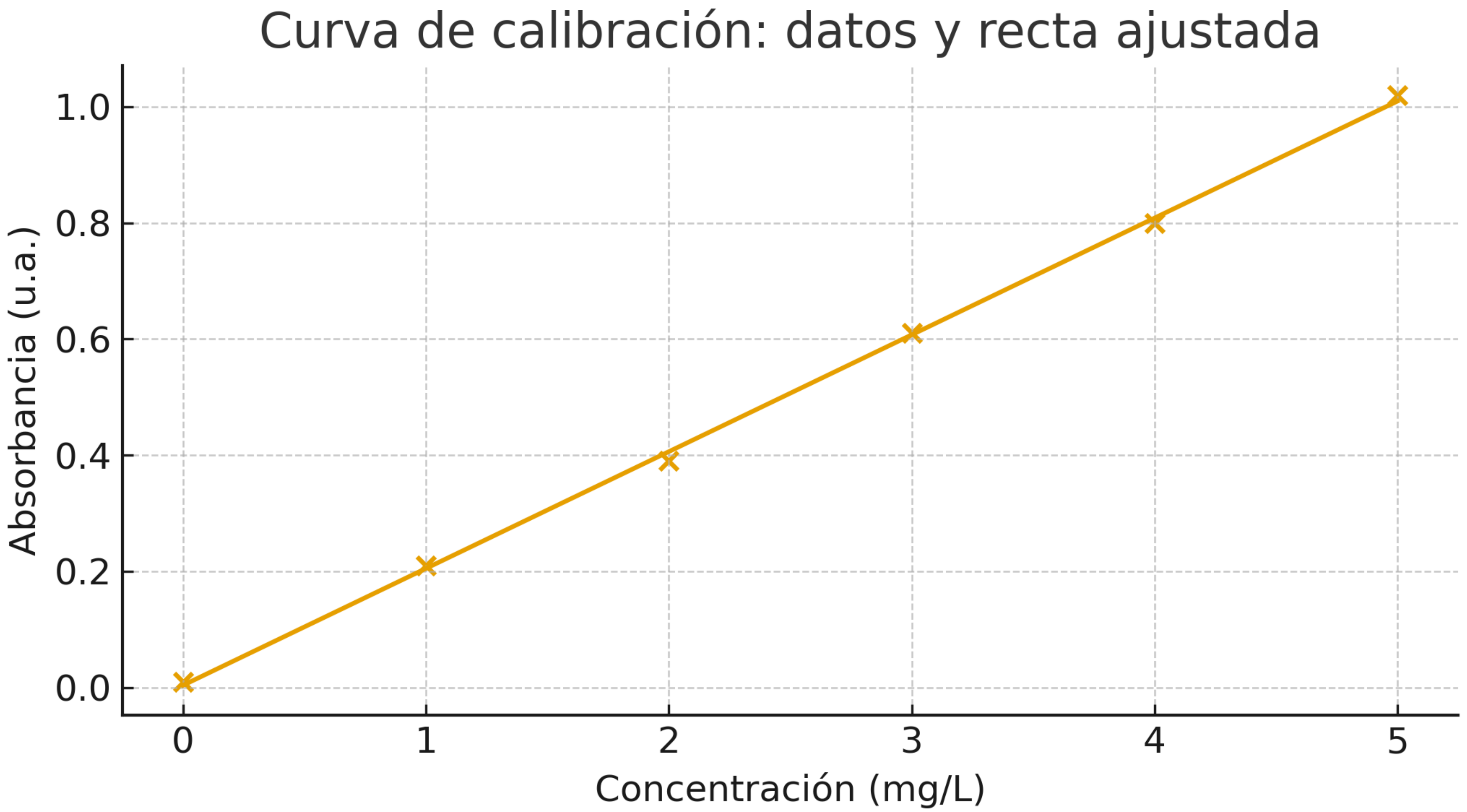
$$IC_{95\%}(\beta_0) = b_0 \pm t_{0.025,4} \cdot SE(b_0)$$

$$IC_{95\%}(\beta_0) = 0.0038 \pm 2.776 \times 0.00818$$

$$IC_{95\%}(\beta_0) \approx (-0.0189,\ 0.0265)$$

✔ Este intervalo **sí incluye 0**, coherente con la prueba t:

el intercepto podría ser 0 en la población.


Curva de calibración: datos y recta ajustada
Absorbancia (u.a.)
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
Concentración (mg/L)

Explorar gráficos de tendencia y linealización de modelos en Excel.

A) Modelo exponencial

$$Y = ae^{bX}$$

Ejemplos:

- cinética de primer orden
- crecimiento bacteriano
- degradación de fármacos

Transformación:

$$\ln Y = \ln a + bX$$

B) Modelo de potencia

$$Y = aX^b$$

Ejemplos:

- procesos biológicos
- leyes de escala
- difusión
- viscosidad vs temperatura

Transformación:

$$\ln Y = \ln a + b \ln X$$

C) Modelo logarítmico

$$Y = a \ln X + b$$

Usado en:

- procesos que crecen rápido al inicio y luego se estabilizan
- química ambiental

No requiere transformación adicional.

D) Modelo cinético de primer orden (muy típico en QFB)

$$\ln A = -kt + \ln A_0$$

Ejemplos:

- degradación de medicamentos
- cinética de reacciones simples
- oxidación en alimentos

Se linealiza graficando:

$$t \text{ vs } \ln A$$

E) Modelo Michaelis–Menten

$$v = \frac{V_{max} S}{K_m + S}$$

Opciones de linealización:

- Lineweaver–Burk
- Hanes–Woolf
- Eadie–Hofstee