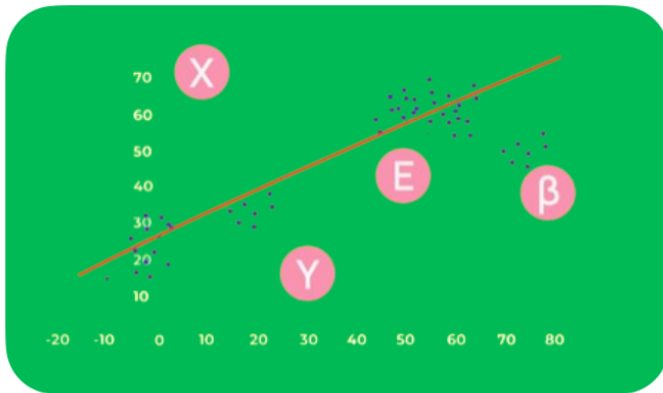


# Regresión Lineal

Consideraciones para el análisis de datos en los cursos de laboratorio de Física



- 1 El problema de la regresión lineal
- 2 La regresión lineal simple
- 3 Método de mínimos cuadrados
- 4 Coeficiente de regresión
- 5 Coeficiente de correlación lineal
- 6 El contraste de regresión
- 7 Inferencias acerca de los parámetros
- 8 Inferencias acerca de la predicción
- 9 Los supuestos del modelo de regresión lineal
- 10 Un ejemplo en donde no se cumplen los supuestos

# 1. El problema de la regresión lineal

- La regresión es una técnica estadística para investigar y modelar relaciones entre variables.
- Las relaciones estadísticas difieren de las funcionales porque no son perfectas; las observaciones no caen directamente sobre una curva.
- Se supone una relación entre una respuesta cuantitativa  $y$  y  $k$  predictores  $x_1, x_2, \dots, x_k$  de la forma general:

$$y = f(x) + \varepsilon$$

donde  $f$  es una función desconocida de  $x_1, \dots, x_k$  y  $\varepsilon$  es un término de error aleatorio independiente de  $x$  con media cero.

- La función  $f$  representa la información sistemática que  $x$  proporciona sobre  $y$ .
- El método paramétrico más utilizado asume que  $f$  es lineal en  $x$ :

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

- Para ajustar el modelo lineal, se estiman los parámetros  $\beta_0, \beta_1, \dots, \beta_k$  de manera que:

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

## 2. La regresión lineal simple

- Modelo:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

donde  $y_i$  es la respuesta,  $x_i$  es el regresor,  $\beta_0$  es la intersección,  $\beta_1$  es la pendiente, y  $\varepsilon_i$  es el término de error aleatorio.

- Características del Modelo:

1.  $y_i$  es una variable aleatoria compuesta por  $\beta_0 + \beta_1 x_i$

2. La función de regresión:

$$E(y_i | x_i) = \beta_0 + \beta_1 x_i$$

relaciona las medias de las distribuciones de  $y_i$  para cada  $x_i$ .

3.  $\varepsilon_i$  introduce variabilidad adicional a  $y_i$  con varianza constante  $\sigma^2$ .

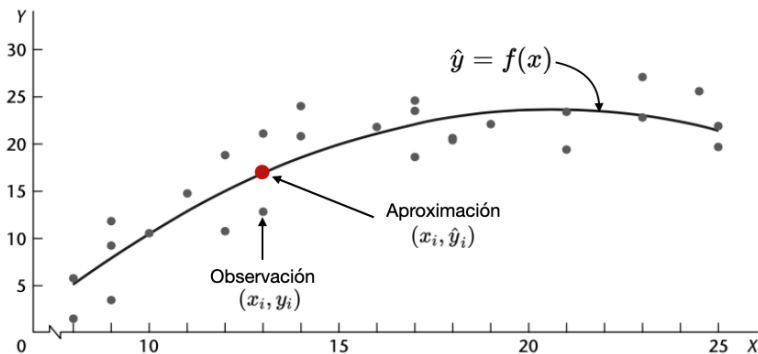
4. El modelo asume varianza constante para  $y_i$ :

$$\sigma^2 \{y_i\} = \sigma^2$$

5. Los términos de error  $\varepsilon_i$  no están correlacionados entre sí, lo que implica que las respuestas  $y_i$  tampoco lo están.

Con técnicas de regresión de una variable  $y$  (respuesta) sobre una variable  $x$  (regresor), se busca una función que sea una buena aproximación de una nube de puntos, mediante una curva del tipo:

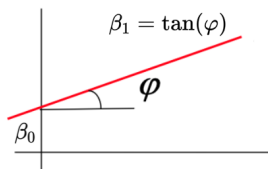
$$\hat{y} = f(x)$$



Un modelo con un único regresor  $x$  y que tiene una relación con la respuesta  $y$  en la forma de una línea recta es lo que se denomina un modelo de regresión lineal simple:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

En donde  $\beta_0$  es la ordenada en el origen (el valor que toma  $y$  cuando  $x$  vale 0),  $\beta_1$  es la pendiente de la recta (e indica cómo cambia  $y$  al incrementar  $x$  en una unidad) y  $\varepsilon$  una variable que incluye un conjunto grande de factores, cada uno de los cuales influye en la respuesta sólo en pequeña magnitud, a la que llamaremos error.



Por lo tanto,  $x$  e  $y$  son variables aleatorias, por lo que no se puede establecer una relación lineal exacta entre ellas.

- Ejemplo: Ley de enfriamiento de Steinhart-Hart.

Para describir con precisión la relación entre la resistencia  $R$  de un termistor y su temperatura  $T$  se utiliza la ecuación de Steinhart-Hart:

$$\frac{1}{T} = A + B \ln(R) + C[\ln(R)]^3$$

donde:

- $T$  es la temperatura en grados Kelvin.
- $R$  es la resistencia en ohmios.
- $A, B, C$  son constantes específicas del termistor.

En un rango pequeño de temperaturas, la relación entre  $R$  y  $T$  puede aproximarse a:

$$R \approx R_0 + \alpha (T - T_0)$$

es decir, a una línea recta donde:  $R_0$  es la resistencia a una temperatura de referencia  $T_0$  y  $\alpha$  es el coeficiente de temperatura, que indica la tasa de cambio de la resistencia con la temperatura.

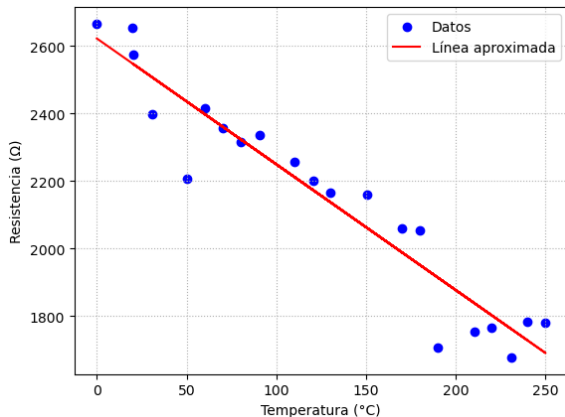
### **Experimento:** Medición de la Resistencia de un Termistor NTC

Descripción del Experimento:

- **Objetivo:** Determinar la relación lineal aproximada entre la resistencia y la temperatura de un termistor NTC en un rango limitado de temperaturas.
- **Variables:**  $y$ , resistencia del material (en ohmios) y  $x$  la temperatura (en grados Kelvin)

## Procedimiento:

- Calentar el Termistor: Coloca el termistor en un baño de agua caliente para llevarlo a una temperatura significativamente más alta que la ambiente.
- Medir la Resistencia y Temperatura: Retira el termistor del agua caliente y mide su resistencia a intervalos regulares mientras se enfría. Simultáneamente, mide la temperatura del termistor.
- Registrar los Datos: Anota las mediciones de resistencia  $R$  y temperatura  $T$





### 3. Método de mínimos cuadrados

El método de los mínimos cuadrados permite estimar  $\beta_0$  y  $\beta_1$  de forma que la suma de los cuadrados de las diferencias entre las observaciones  $y_i$  y la recta sea un mínimo.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (1)$$

La ecuación (1) es un modelo de regresión muestral, escrito en términos de los  $n$  pares de datos  $(y_i, x_i)$ , con  $i = 1, 2, \dots, n$ .

A los estimadores obtenidos por mínimos cuadrados  $\beta_0$  y  $\beta_1$ , los llamaremos  $b$  y  $m$ , respectivamente, y deben satisfacer una relación lineal de la forma:

$$\hat{y} = mx + b,$$

donde  $\hat{y}$  es la variable dependiente y  $x$  es la variable independiente, en nuestro caso la magnitud controlada por el experimentador. El método de mínimos cuadrados consiste en minimizar suma de los cuadrados de los errores:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Es decir, la suma de los cuadrados de las diferencias entre los valores reales observados ( $y_i$ ) y los valores estimados ( $\hat{y}_i$ ).

Minimizar suma de los cuadrados de los errores se logra calculando las derivadas parciales de la suma con respecto a  $m$  y con respecto a  $b$ , e igualándolas a cero.

Con este método, las expresiones que se obtiene para  $b$  y  $m$  son las siguientes:

$$m = \frac{S_{xy}}{S_x^2} \quad b = \bar{y} - m\bar{x},$$

En donde  $\bar{x}$  e  $\bar{y}$  denotan las medias muestrales de  $x$  y  $y$  (respectivamente),

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \quad \bar{y} = \frac{\sum_{i=1}^n y_i}{n},$$

$S_x^2$  es la varianza muestral de  $x$  y  $S_{xy}$  es la covarianza muestral entre  $x$  y  $y$ . Estos parámetros se calculan como:

$$S_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}, \quad S_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}, \quad S_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}.$$

La cantidad  $m$  se denomina coeficiente de regresión de  $y$  sobre  $x$ , lo denotamos por  $m_{y/x}$ .

En nuestro ejemplo, los estadísticos descriptivos anteriores para las variables temperatura y tiempo del enfriamiento son los siguientes:

$$\bar{x} = 125,0119, \quad \bar{y} = 2156,81$$

$$S_x^2 = 12896,0685, \quad S_y^2 = 196582,6099$$

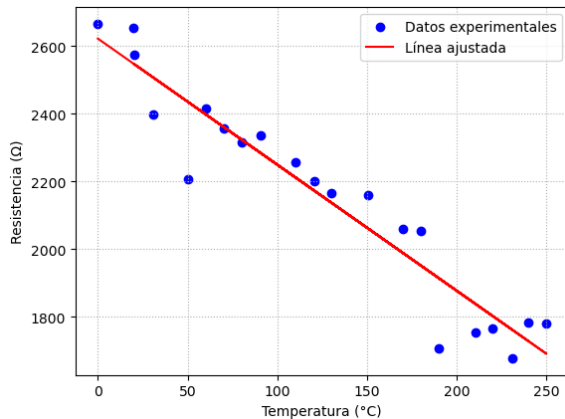
$$S_{xy} = -48017,6465$$

$$m = -3,7234, \quad b = 2622,2834$$

La recta de regresión ajustada es la siguiente:

$$\hat{y} = 2622,2834 - 3,7234x,$$

donde  $\hat{y}$  es la resistencia y  $x$  la temperatura.



Hay varias propiedades que se cumplen para los mínimos cuadrados:

- ① La suma de los valores observados  $y_i$  es igual a la suma de los valores ajustados  $\hat{y}_i$ , o bien

$$\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$$

- ② La suma de los residuos que contiene un intercepto  $\beta_0$  es siempre cero

$$\sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n e_i = 0$$

- ③ La recta siempre pasa por el centroide, el punto  $(\bar{x}, \bar{y})$  de los datos.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = m\bar{x} + b.$$

- ④ La suma de los residuos ponderada por el valor de la variable regresora es cero

$$\sum_{i=1}^n x_i e_i = 0$$

- ⑤ La suma de los residuos ponderada por el valor ajustado siempre es cero

$$\sum_{i=1}^n \hat{y}_i e_i = 0$$

## 4. Coeficiente de regresión

El coeficiente de regresión nos da información sobre el comportamiento de la variable  $y$  frente a la variable  $x$ , de manera que:

- ① Si  $m_{y/x} = 0$ , para cualquier valor de  $x$  la variable  $y$  es constante.
- ② Si  $m_{y/x} > 0$ , esto nos indica que al aumentar el valor de  $x$ , también aumenta el valor de  $y$ .
- ③ Si  $m_{y/x} < 0$ , esto nos indica que al aumentar el valor de  $x$ , el valor de  $y$  disminuye.

En el ajuste de regresión lineal para la resistencia y la temperatura resultó en

$$\hat{y} = 2622,2834 - 3,7234x,$$

El coeficiente de regresión que se obtuvo fue  $m_{y/x} = -3,7234 < 0$  y esto indica que al aumentar  $x$  disminuye  $y$ .

## 5. Coeficiente de correlación lineal

El coeficiente de correlación lineal entre  $x$  e  $y$  viene dado por:

$$r = \frac{S_{xy}}{S_x S_y}$$

y trata de medir la dependencia lineal que existe entre las dos variables. Su cuadrado se denomina coeficiente de determinación  $r^2$ .

Propiedades del coeficiente de correlación:

- 1 No tiene dimensión, y siempre toma valores en  $[-1,1]$ .
- 2 Si las variables son independientes, entonces  $r = 0$ , el inverso no tiene por qué ser cierto.
- 3 Si existe una relación lineal exacta entre  $x$  e  $y$ , entonces  $r = 1$  (relación directa) ó  $r = -1$  (relación inversa).
- 4 Si  $r > 0$ , indica una relación directa entre las variables (si aumenta  $x$ , aumenta  $y$ ).
- 5 Si  $r < 0$ , indica una relación inversa entre las variables (si aumenta  $x$ , disminuye  $y$ ).

Para nuestro ejemplo el valor de  $r$  es

$$r = \frac{S_{xy}}{S_x S_y} = \frac{-48017,6465}{\sqrt{12896,0685} \sqrt{196582,6099}} = -0,9537$$

El menos indica que existe una relación inversa entre las variables. Además su valor es próximo a 1 indicando una dependencia lineal muy fuerte.

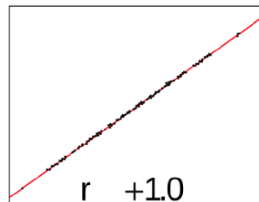
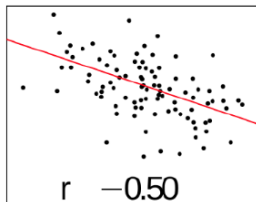
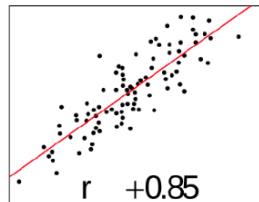
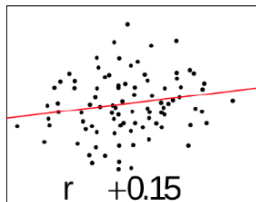
El coeficiente de determinación al cuadrado es  $r^2 = 0,9095$ .

La relación entre los coeficientes de regresión y de correlación:

$$m_{y/x} = r \frac{S_y}{S_x} = -3,7234,$$

$$m_{x/y} = r \frac{S_x}{S_y} = -0,2443$$

Los dos coeficientes de regresión y el coeficiente de correlación tienen el mismo signo















OLS Regression Results						
=====						
Dep. Variable:	y	R-squared:	0.963			
Model:	OLS	Adj. R-squared:	0.958			
Method:	Least Squares	F-statistic:	205.7			
Date:	Mon, 27 May 2024	Prob (F-statistic):	5.45e-07			
Time:	10:18:42	Log-Likelihood:	-3.7692			
No. Observations:	10	AIC:	11.54			
Df Residuals:	8	BIC:	12.14			
Df Model:	1					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
const	0.7362	0.226	3.254	0.012	0.215	1.258
x1	5.8844	0.410	14.343	0.000	4.938	6.831
=====						
Omnibus:	1.614	Durbin-Watson:		1.076		
Prob(Omnibus):	0.446	Jarque-Bera (JB):		0.806		
Skew:	0.218	Prob(JB):		0.668		
Kurtosis:	1.679	Cond. No.		4.04		
=====						