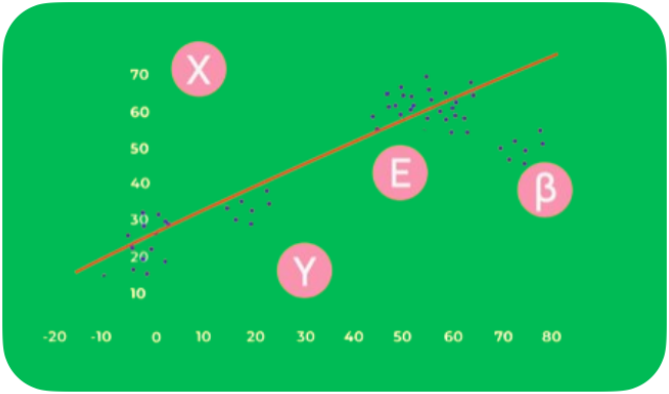
Universidad Industrial de Santander

Regresión Lineal

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



Regresión Lineal

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 $x_1, x_2, \ldots, x_k$ de la forma general:

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

  donde f es una función desconocida de $x_1, \ldots, x_k$ y ε 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:

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

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

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

- Modelo:

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

donde y_i es la respuesta, x_i es el regresor, β_0 es la intersección, β_1 es la pendiente, y ε_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:

$$E\left(y_i \mid x_i\right) = \beta_0 + \beta_1 x_i$$

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

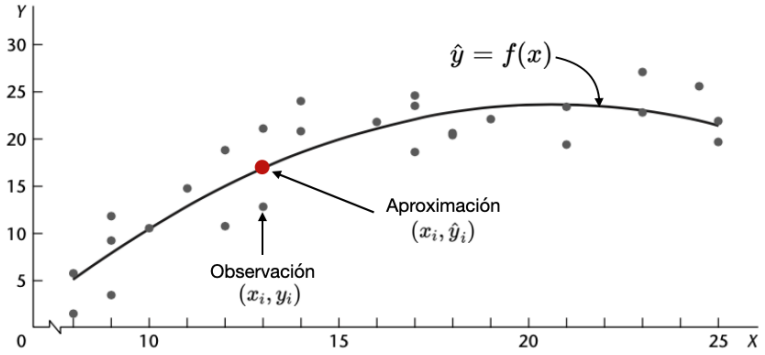
3. ε_i introduce variabilidad adicional a y_i con varianza constante σ^2.
4. El modelo asume varianza constante para y_i:

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

5. Los términos de error ε_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:

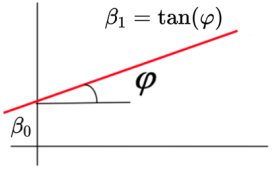
$$\hat{y} = f(x)$$



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 β_0 es la ordenada en el origen (el valor que toma y cuando x vale 0), β_1 es la pendiente de la recta (e indica cómo cambia y al incrementar x en una unidad) y ε 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 Newton.

  Esta ley establece que la tasa de cambio de la temperatura de un objeto es proporcional a la diferencia entre la temperatura del objeto y la temperatura del ambiente:

$$\frac{dT}{dt} = -k\,(T - T_{\mathsf{e}})$$

  donde:
  - T es la temperatura del objeto.
  - T_{e} es la temperatura del entorno.
  - k es una constante que depende de las características del objeto y del entorno.
  - $\frac{dT}{dt}$ es la tasa de cambio de la temperatura con respecto al tiempo.

La solución de esta ecuación diferencial es:

$$T(t) = T_{\mathsf{e}} + (T_0 - T_{\mathsf{e}})\,e^{-kt}$$

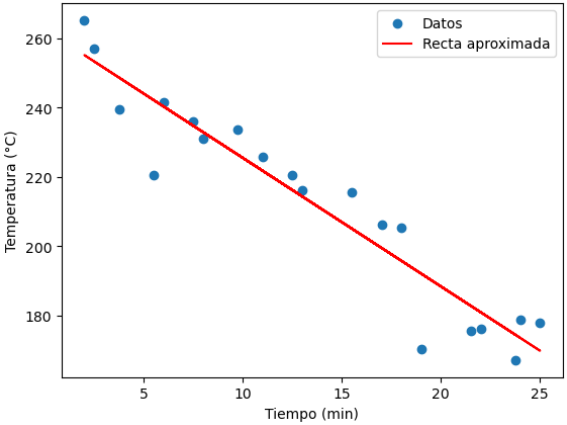
donde: $T(t)$ es la temperatura del objeto en el tiempo t y T_0 es la temperatura inicial.

Experimento: Enfriamiento de un material con exposición a un medio refrigerante
Descripción del Experimento:

- Objetivo: Determinar cómo la temperatura de un material disminuye con el tiempo cuando se expone a un medio refrigerante.
- Variables: y, temperatura del material (en grados Celsius) y x tiempo de exposición al refrigerante (en minutos)

Procedimiento:

- Inicialización: Calentar un material a una temperatura inicial alta.
- Exposición: Colocar el material en un medio refrigerante.
- Medición: Registrar la temperatura del material a intervalos regulares de tiempo
- Datos Recogidos:
  - y_n : Temperatura del material en diferentes momentos.
  - x_n : Tiempo transcurrido desde el inicio del enfriamiento.



3. Método de mínimos cuadrados

El método de los mínimos cuadrados permite estimar β_0 y β_1 de forma que la suma de los cuadrados de las diferencias entre las observaciones y_i y la recta sea un mínimo.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{1}$$

La ecuación (1) es un modelo de regresión muestral, escrito en términos de los n pares de datos (y_i, x_i), con $i = 1, 2, \ldots, n$.
A los estimadores obtenidos por mínimos cuadrados β_0 y β_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),

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

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:

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

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

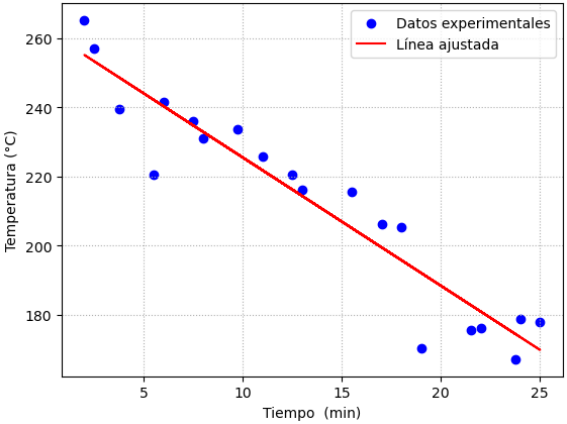
En nuestro ejemplo, los estadísticos descriptivos anteriores para las variables temperatura y tiempo del enfriamiento son los siguientes:

$$\bar{x} = 13{,}3625, \quad \bar{y} = 213{,}0075$$
$$S_x^2 = 110{,}6559, \quad S_y^2 = 1692{,}2956$$
$$S_{xy} = -410{,}3117$$
$$m = -3{,}708, \quad b = 262{,}556$$

La recta de regresión ajustada es la siguiente:

$$\hat{y} = 262{,}556 - 3{,}708x\,,$$

donde $\hat{y}$ es la temperatura y x el tiempo de enfriamiento.



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 β_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 temperatura y el tiempo de enfriamiento resultó

$$\hat{y} = 262{,}556 - 3{,}708x\,,$$

El coeficiente de regresión que se obtuvo fue $m_{y/x} = -3{,}708 < 0$ 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 directa entre las variables (si aumenta x, disminuye y).

Para nuestro ejemplo el valor de r es

$$r = \frac{S_{xy}}{S_x S_y} = \frac{-410,3117}{\sqrt{110,6559}\sqrt{1692,2956}} = -0,9482$$

Al ser negativo, esto indica que existe una relación inversa entre las variables cizallamiento y edad del pegamento. Además su valor es próximo a 1 indicando una dependencia lineal muy fuerte.

El coeficiente de determinación al cuadrado es $r^2 = 0,8990$.

Relación entre los coeficientes de regresión y de correlación:

$$m_{y/x} = r\frac{S_y}{S_x} = -3,708, \quad m_{x/y} = r\frac{S_x}{S_y} = -0,242.$$

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

Gráficos

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