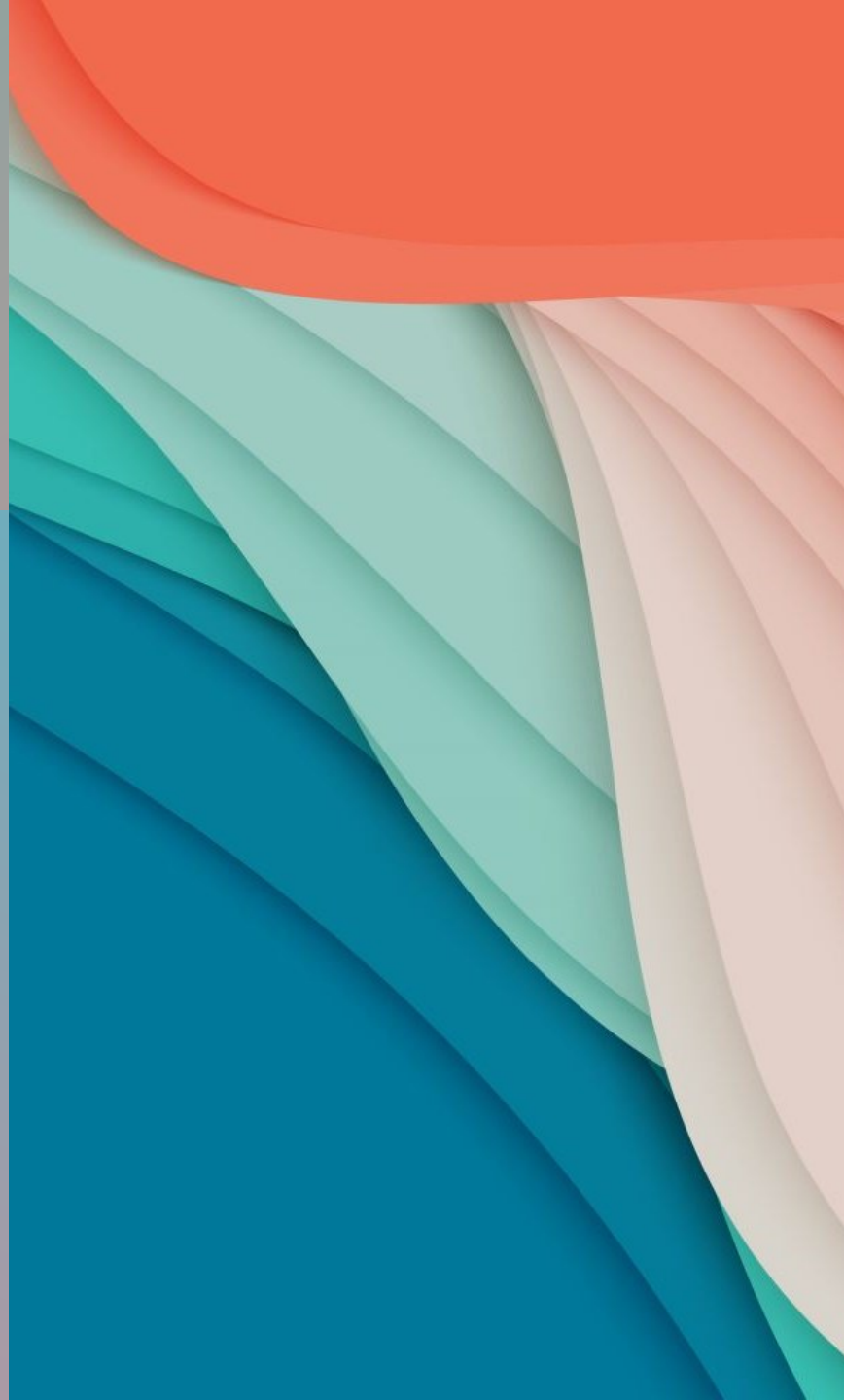
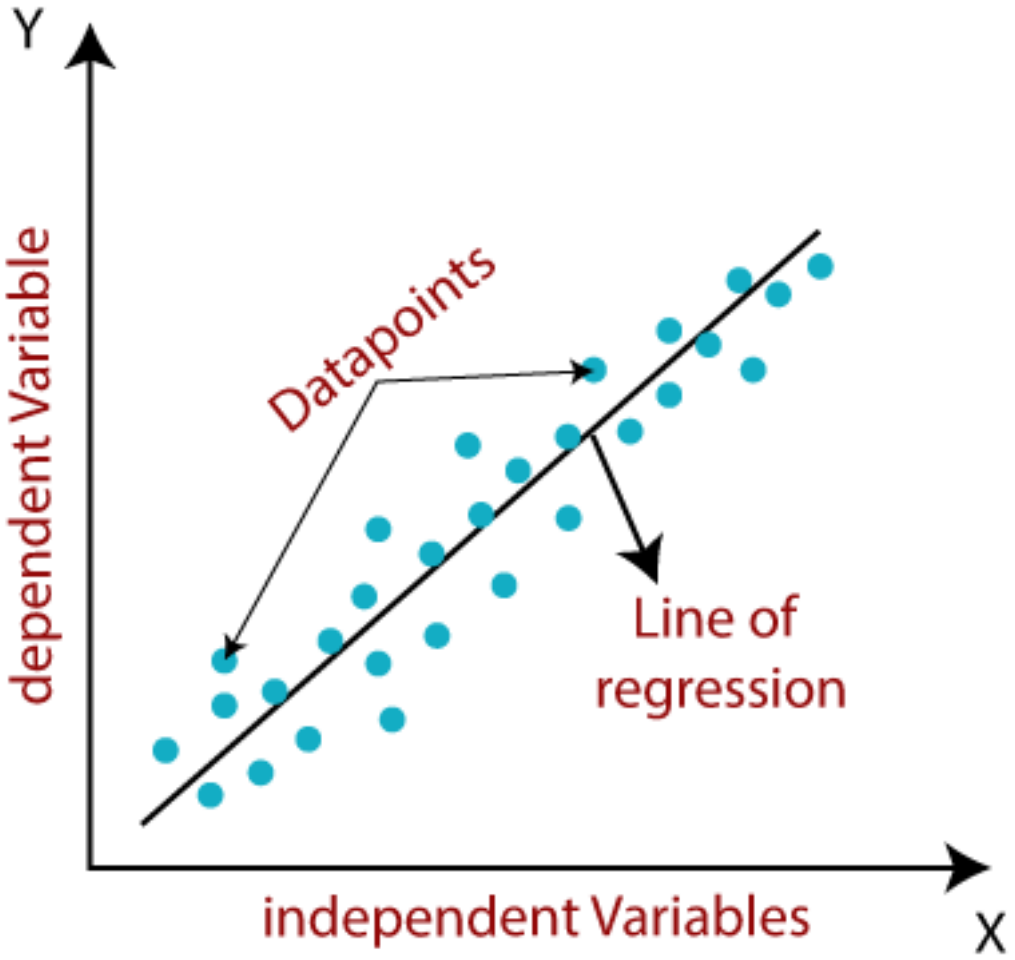
CIENCIA DE DATOS PARA FÍSICOS

REGRESIÓN LINEAL



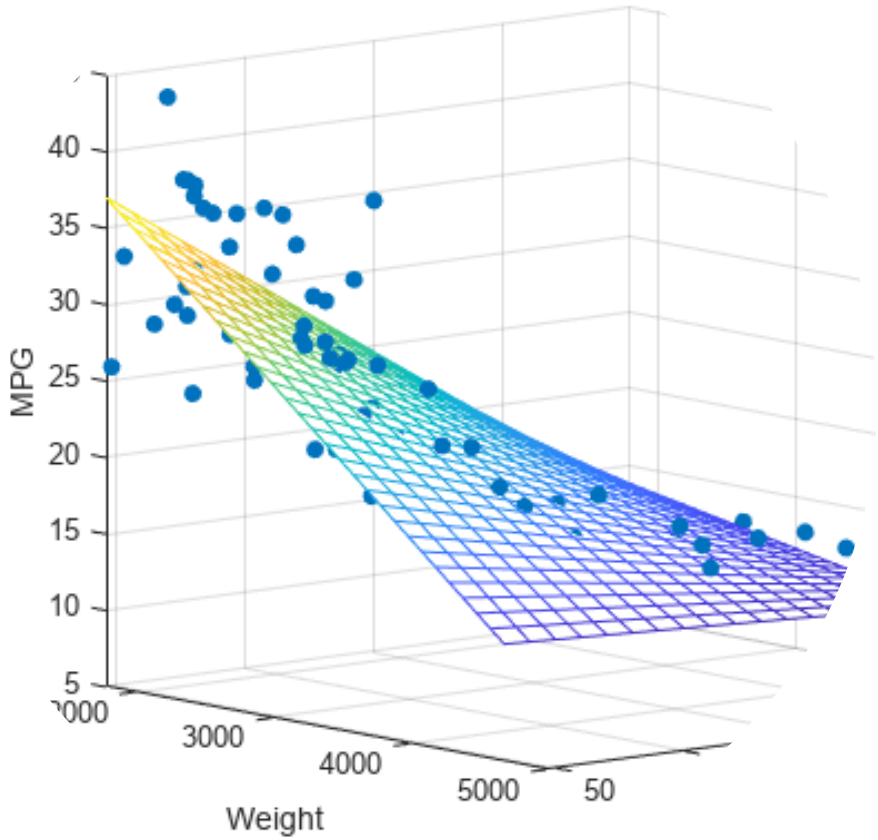
INTRODUCCIÓN

La regresión lineal es una técnica de ***aprendizaje automático supervisado*** que se utiliza para predecir una variable de salida continua (y) a partir de una o varias variables de entrada (x). Es una de las técnicas de análisis estadístico más utilizadas y una de las más sencillas de entender.



TIPOS DE REGRESIÓN LINEAL

Existen dos tipos de regresión lineal: la regresión lineal ***simple*** y la regresión lineal ***múltiple***. En la regresión lineal simple, se utiliza una única variable de entrada (x) para predecir la variable de salida (y). En la regresión lineal múltiple, se utilizan varias variables de entrada (x1, x2, ..., xn) para predecir la variable de salida (y).



DELO DE ESIÓN INEAL

El modelo de regresión lineal se expresa como $y = b0 + b1x + \varepsilon$, donde y es la variable de salida, x es la variable de entrada, $b0$ y $b1$ son los coeficientes de regresión y ε es el término de error. El objetivo de la regresión lineal es encontrar los valores de $b0$ y $b1$ que minimicen la suma de los errores al cuadrado.

$$f_\theta(\vec{\theta}, \vec{x}) = \vec{\theta}^T \vec{x} = \sum_{i=0}^{n} \theta_i x_i$$

Asumiendo ruido épsilon con distribución normal en el modelo:

$$f_\theta(\vec{\theta}, \vec{x}) + \varepsilon$$

Se modela la probabilidad condicional de y, dados los datos y los parámetros

$$P(y|x, \theta, \sigma^2) = N(f_\theta\ (\vec{\theta}, \vec{x}), \sigma^2)$$

MÉTODO DE MÍNIMOS CUADRADOS

El método de mínimos cuadrados es una técnica utilizada para encontrar los valores de b0 y b1 que minimizan la suma de los errores al cuadrado. Este método consiste en calcular la suma de los errores al cuadrado para cada valor de b0 y b1 y elegir los valores que dan como resultado la menor suma de errores al cuadrado.

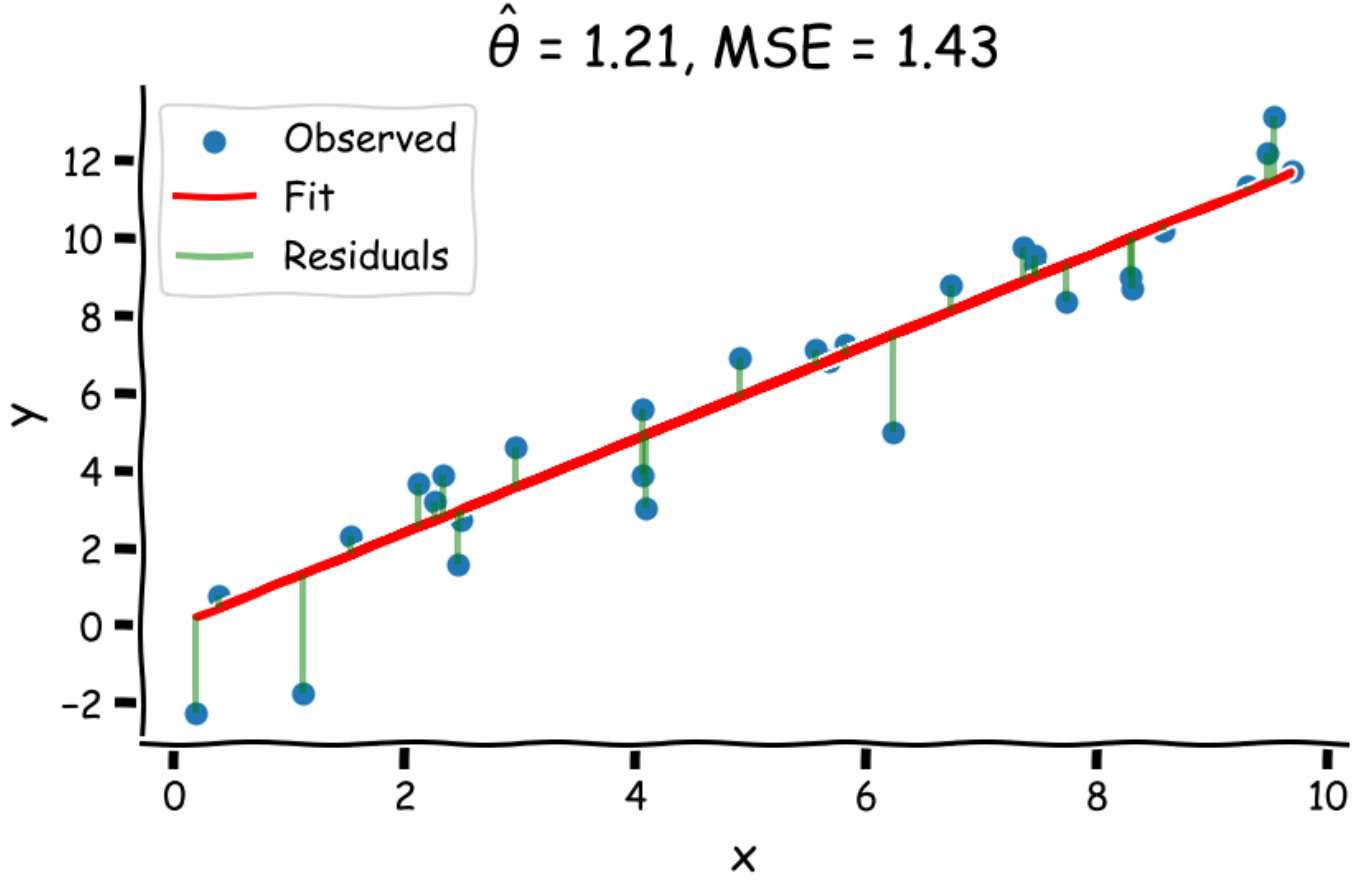
$$m = \frac{n \cdot \Sigma(x \cdot y) - \Sigma x \cdot \Sigma y}{n \cdot \Sigma x^2 - |\Sigma x|^2}$$

$$b = \frac{\Sigma y \cdot \Sigma x^2 - \Sigma x \cdot \Sigma(x \cdot y)}{n \cdot \Sigma x^2 - |\Sigma x|^2}$$

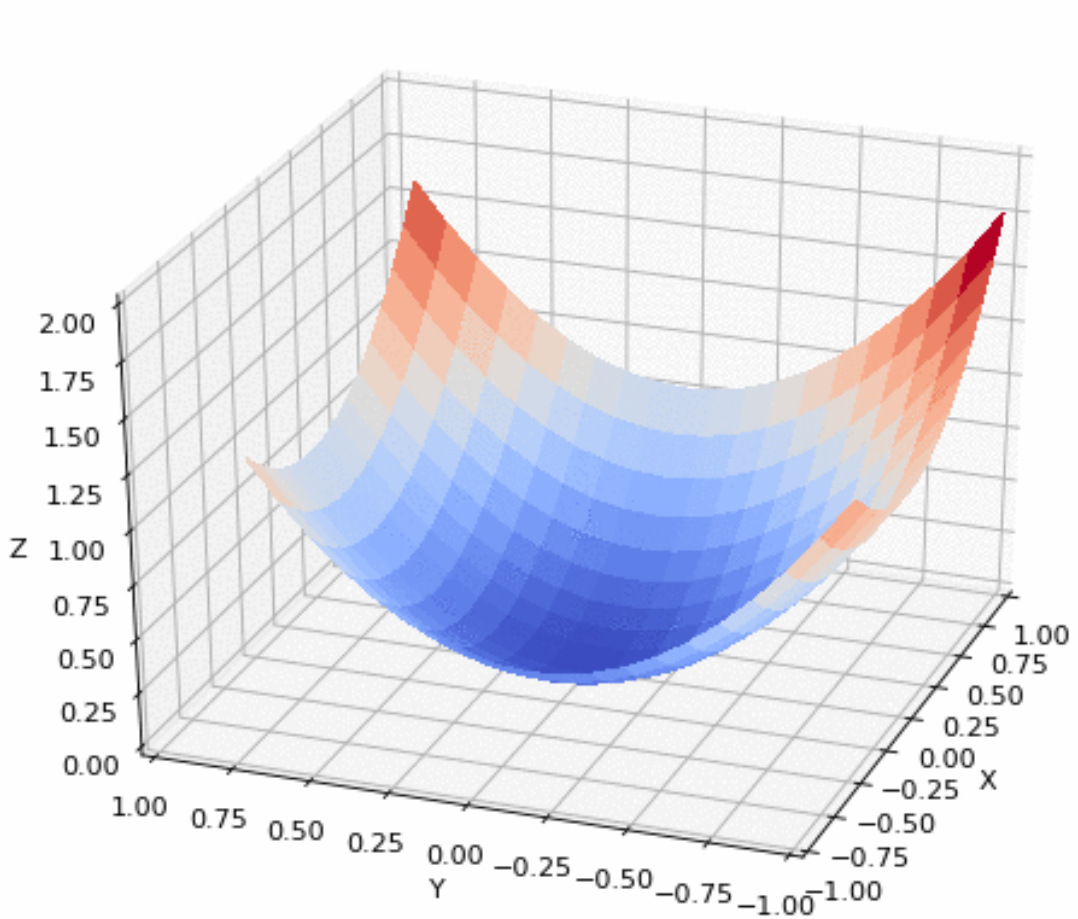
¿QUÉ FUNCIONES MINIMIZAMOS?

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$



¿QUÉ FUNCIONES MINIMIZAMOS?



¿CÓMO LO HACEMOS?

DESCENSO DEL GRADIENTE

Es un algoritmo iterativo de primer orden que actualiza todos los parámetros a modo de que el Error se minimice de forma local.

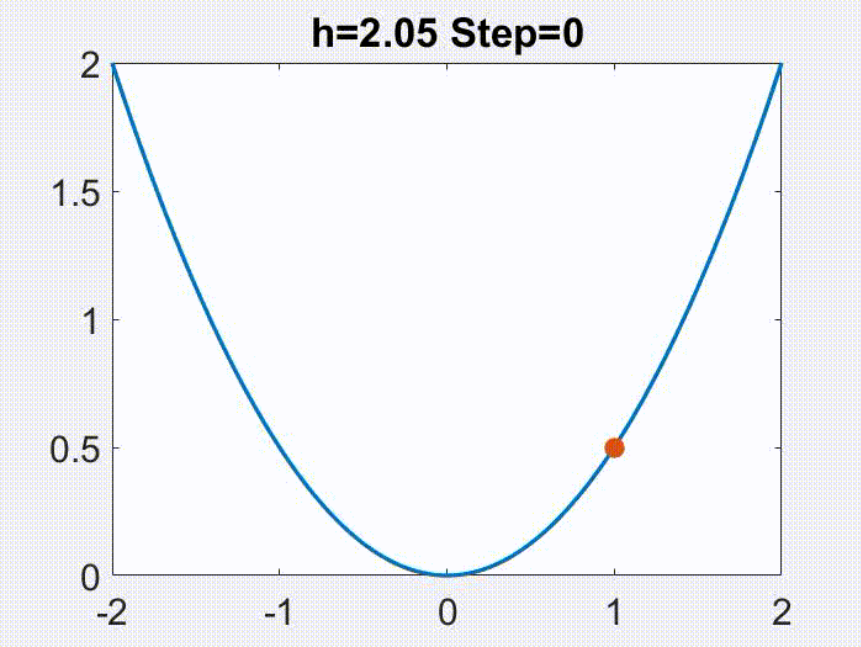
$$\boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^{[t]} - \alpha \nabla E(\boldsymbol{\theta}^{[t]})$$

$$\nabla E(\boldsymbol{\theta}^{[t]}) = \left[\frac{\partial E}{\partial \theta_0^{[t]}}, \cdots, \frac{\partial E}{\partial \theta_d^{[t]}} \right]$$

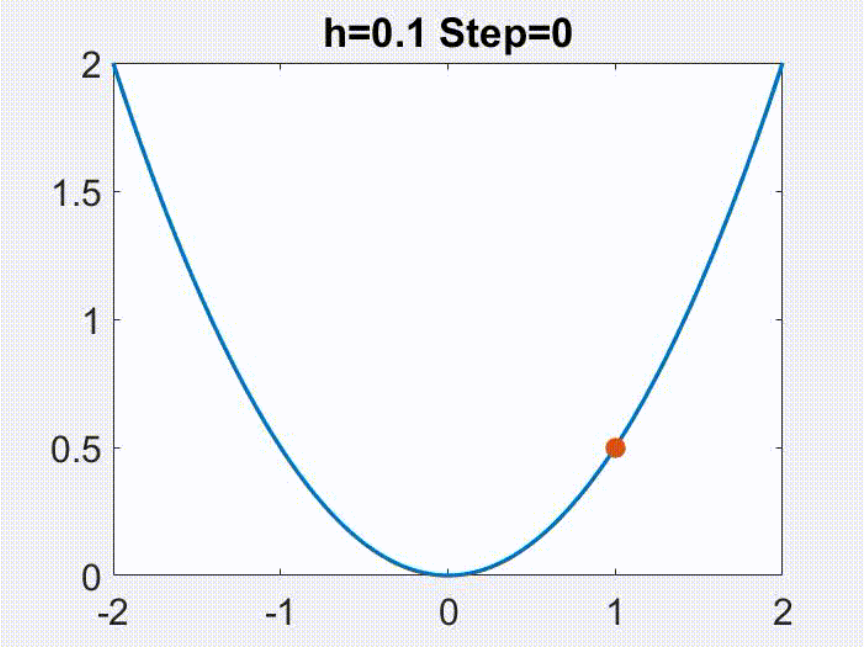
α Se le conoce como tasa de aprendizaje, el cual es un hiperparámetro, es decir, el modelo no lo aprende, si no que hay que darle un valor a priori.

A cada iteración dada sobre todo el dataset se le conoce como época

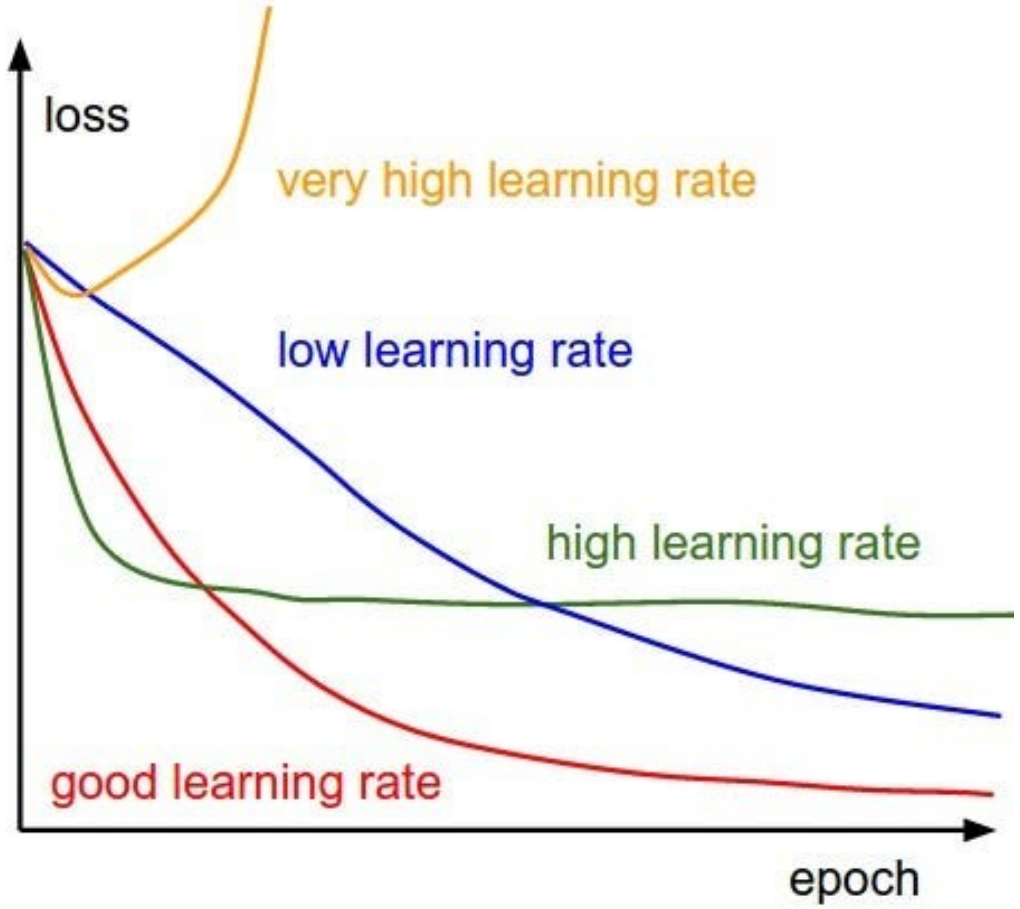
DESCENSO DEL GRADIENTE



Tasa de aprendizaje grande



Tasa de aprendizaje pequeña



DESCENSO DEL GRADIENTE

COEFICIENTE DE DETERMINACIÓN (R^2)

El coeficiente de determinación (R^2) es una medida de la calidad del modelo de regresión. R^2 se calcula como la proporción de la varianza de y que es explicada por el modelo de regresión. Un valor de R^2 cercano a 1 indica que el modelo es muy bueno en la predicción de la variable de salida, mientras que un valor cercano a 0 indica que el modelo no es muy bueno.

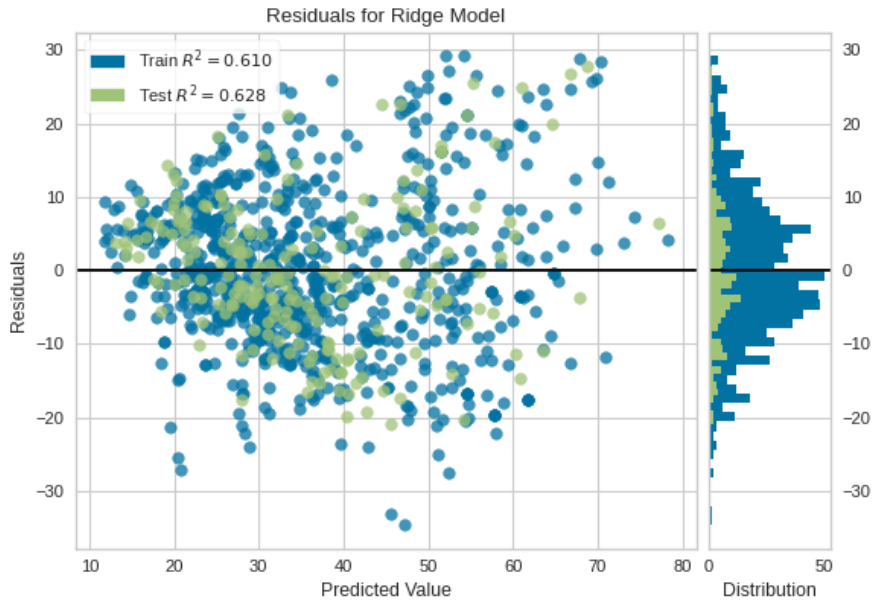
$$R^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} = \rho^2$$

Donde:

- σ_{XY} es la covarianza de (X, Y)
- σ_X^2 es la varianza de la variable X
- σ_Y^2 es la varianza de la variable Y

ANÁLISIS DE RESIDUOS

El análisis de residuos es una técnica utilizada para evaluar la calidad del modelo de regresión. Los residuos son las diferencias entre los valores observados de y y los valores predichos por el modelo. Si los residuos tienen una distribución normal alrededor de cero, entonces se puede concluir que el modelo de regresión es adecuado para los datos.



* Residual plot

SUPUESTOS DE LA REGRESIÓN LINEAL

1. ***Relación Lineal***: Existe una relación lineal entre la(s) variable(s) independiente(s) x y la variable dependiente y (scatter plot)
2. ***Independencia***: Los valores residuales son independientes. En particular no existen correlaciones entre residuales consecutivos en una serie de tiempo.
3. ***Homocedasticidad***: Los residuales tienen varianza constante para cada x. (Fitted Values vs Residuals, debe ser como una función lineal o constante)
4. ***Normalidad***: Los residuales del modelo se distribuyen en una guassiana (Quantile-Quantile Plots, deben ser lineales)

* Si no se cumple alguna de estas, no se debe usar un modelo de regresión lineal para modelar los datos. En algunas ocasiones puede(n) cumplirse los supuestos a través de alguna transformación a los datos

ESCALADO DE CARACTERÍSTICAS

El descenso del gradiente es una excelente técnica para aproximarse al valor mínimo del Error. Sin embargo, si los valores de las variables dependientes son de magnitudes muy diferentes, éste podría fallar debido a dicha diferencia de magnitud.

Para evitar dicho fallo debemos normalizar los rangos a modo de que la contribución de las características sea proporcional.

Existen al menos tres métodos de normalizacion:

1. Re-escalado
2. Estandarización
3. Magnitud unitaria

ESCALADO DE CARACTERÍSTICAS

$$x' = \frac{x - min(x_{1:n})}{max(x_{1:n}) - min(x_{1:n})} \quad \text{(Re-escalado)}$$

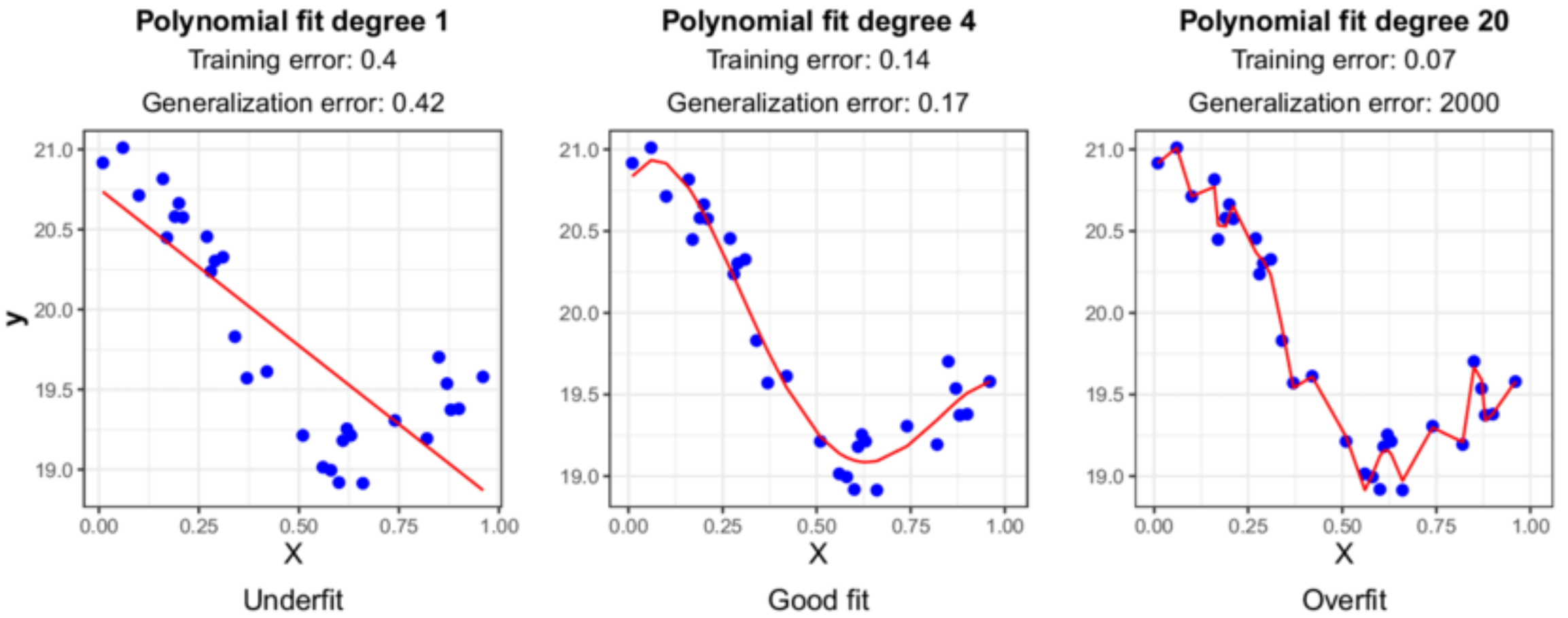
$$x' = \frac{x - \bar{x}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \bar{x})^2}} \quad \text{(Estandaziración)}$$

$$x' = \frac{x}{\|x\|} \quad \text{(Magnitud unitaria)}$$

ESCALADO DE CARACTERÍSTICAS

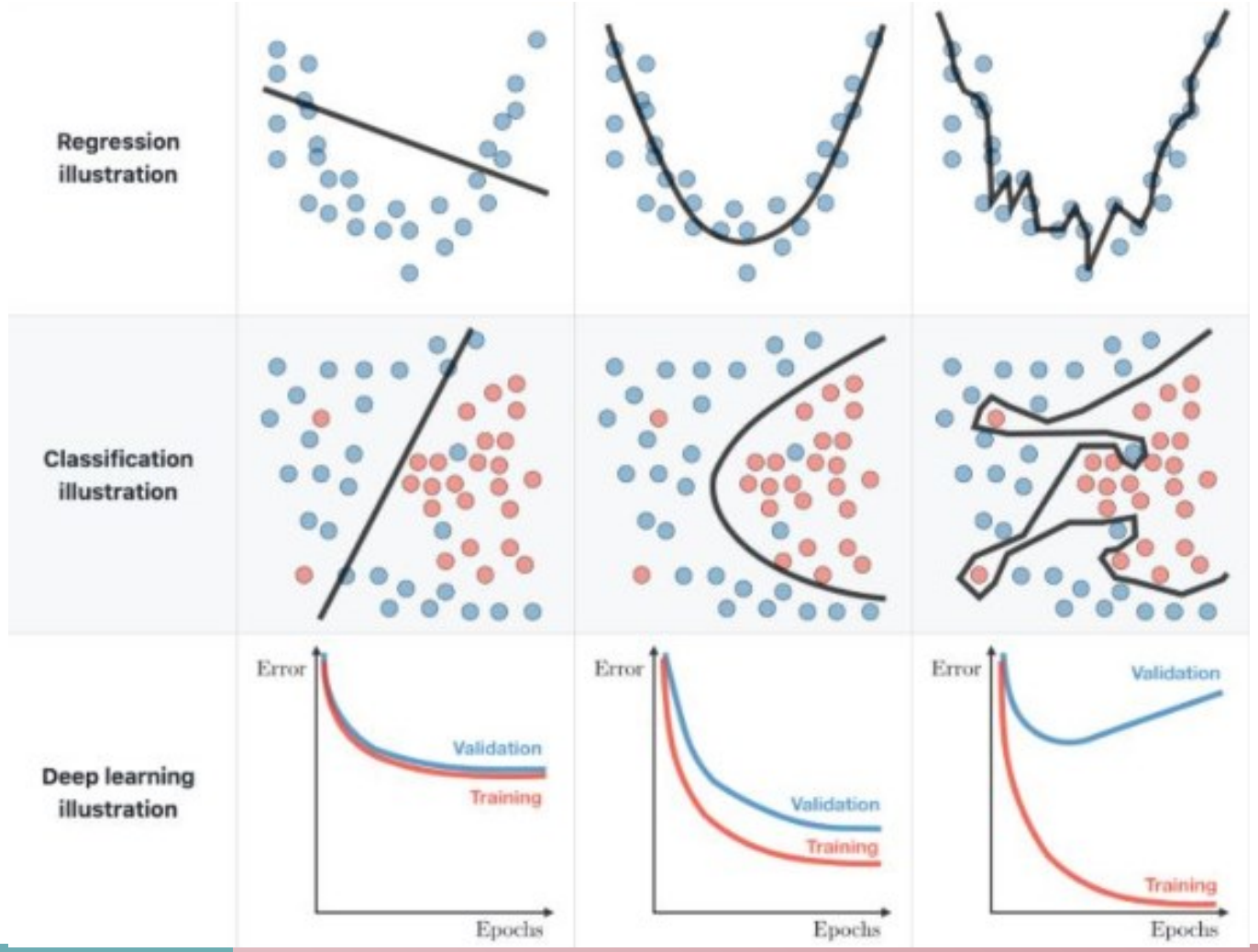
Pro tip: La normalización de rangos puede introducir sesgo al modelo si ésta se aplica sobre todos los datos (como uno solo).

Entonces, asegurate primero de dividir el dataset en entrenamiento-prueba-validación y a cada uno de ellos le aplicas la técnica de normalización ***por separado.***



¿SOBREAJUSTE?

¿SOBREAJUSTE?



REGULARIZACIÓN

La regularización es una técnica utilizada en el aprendizaje automático para ***prevenir el sobreajuste de un modelo***. El sobreajuste ocurre cuando un modelo se ajusta demasiado bien a los datos de entrenamiento y, como resultado, tiene una mala capacidad para generalizar y predecir nuevos datos.

La regularización agrega una penalización a los parámetros del modelo que pueden tomar valores extremos, lo que reduce la complejidad del modelo y lo hace menos propenso a sobreajustarse.

Hay dos tipos principales de regularización:

- La regularización L1
- La regularización L2

REGULARIZACIÓN L1

También conocida como regresión LASSO (Least Absolute Shrinkage and Selection Operator), agrega una penalización a la suma de los valores absolutos de los coeficientes del modelo. Esto significa que los coeficientes del modelo pueden tomar valores cercanos a cero, lo que resulta en la eliminación de características no importantes.

$$+\lambda \sum_{j=0}^{p} |w_j|$$

La cantidad de penalización se controla mediante un hiperparámetro denominado parámetro de regularización. Cuanto mayor sea el valor del parámetro de regularización, mayor será la penalización y menor será la complejidad del modelo.

REGULARIZACIÓN L2

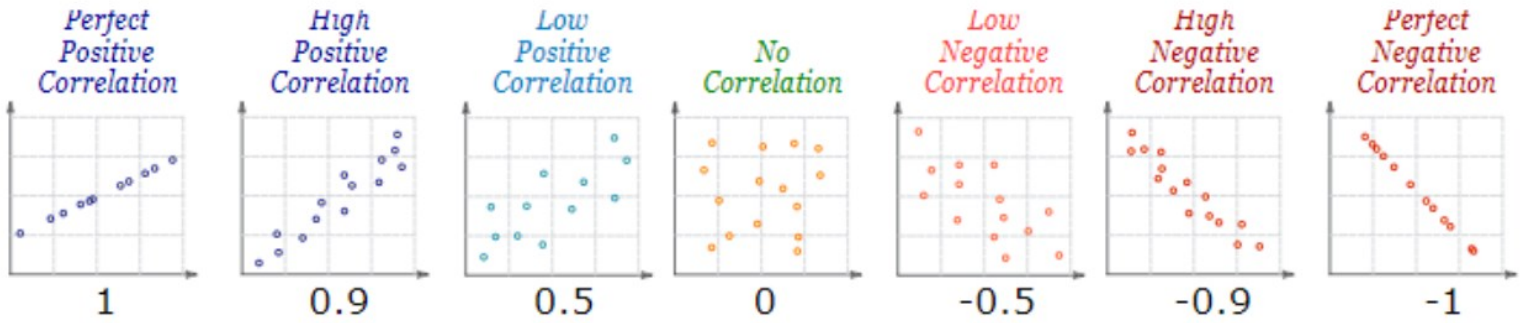
Conocida como regresión de Ridge, agrega una penalización a la suma de los valores cuadrados de los coeficientes del modelo. Esto significa que los coeficientes del modelo no pueden tomar valores extremos y deben estar cerca de cero, lo que reduce la complejidad del modelo.

$$+\lambda \sum_{j=0}^{p} |w_j|^2$$

La cantidad de penalización se controla mediante un hiperparámetro denominado parámetro de regularización. Cuanto mayor sea el valor del parámetro de regularización, mayor será la penalización y menor será la complejidad del modelo.

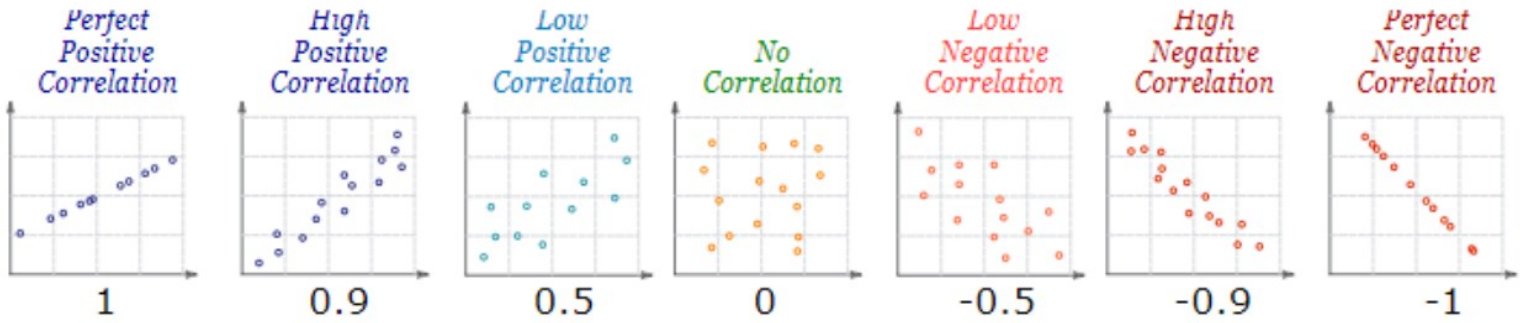
CORRELACIÓN

Medida estadística que describe la relación entre dos variables. Se refiere a la forma en que dos variables cambian juntas. Si dos variables están correlacionadas, significa que cuando una variable cambia, la otra también tiende a cambiar de una manera consistente



CAUSALIDAD

Se refiere a una relación de causa y efecto entre dos variables. Implica que un cambio en una variable causa directamente un cambio en la otra variable. En una relación causal, una variable es la causa y la otra es el efecto.



CORRELACIÓN NO IMPLICA CAUSALIDAD

Aunque dos variables pueden estar correlacionadas, esto no implica necesariamente que exista una relación causal entre ellas.

La falta de causalidad en una correlación puede deberse a varias razones. En primer lugar, puede haber una coincidencia casual o una relación espuria, donde dos variables parecen estar relacionadas, pero en realidad están influenciadas por un tercer factor no considerado.