Regresión Lineal

Sesión Sincrónica

{desafío} latam_

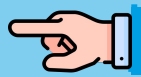
Generar modelos predictivos utilizando scikit-learn de acuerdo a requerimientos

- **Unidad 1: Estadística descriptiva y probabilidades**
  (Parte I)

  (Parte II)
- **Unidad 2: Variable aleatoria**
  (Parte I)

  (Parte II)
- **Unidad 3:Estadística inferencial**
- **Unidad 4: Regresión**
  (Parte I)

  (Parte II)



Te encuentras aquí

{desafío} latam_

¿Qué aprenderás en esta sesión?

Aprenderás sobre los fundamentos y la aplicación de la regresión lineal utilizando Python.

¿Qué entendemos por correlación?

¿Cómo podríamos expresar correlaciones?

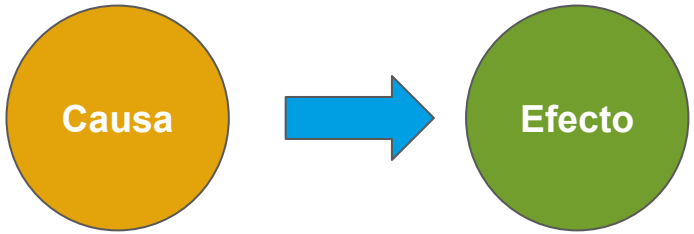
/*Causalidad y correlación*/

Causalidad

Concepto

En el contexto del análisis de datos y la estadística, la causalidad implica que un cambio en una variable (llamada **causa**) produce un cambio directo y medible en otra variable (llamada **efecto**).

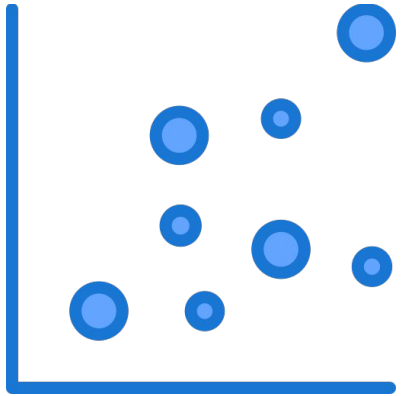
En otras palabras, cuando una variable causa un efecto en otra, se puede establecer una conexión directa y específica entre ambas.



Correlación

Definición

La correlación es una medida estadística que evalúa la relación y la fuerza de asociación entre dos o más variables. Se utiliza para determinar si existe una relación entre las variables, y para medir la dirección y la intensidad de esta relación.



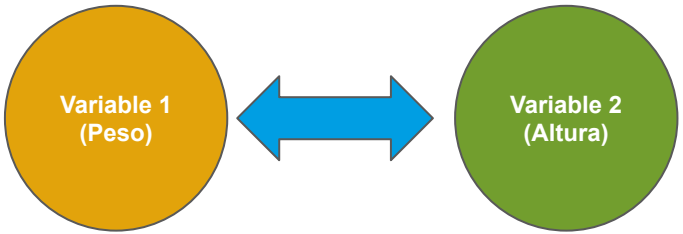
{desafío} latam_

Correlación y causalidad

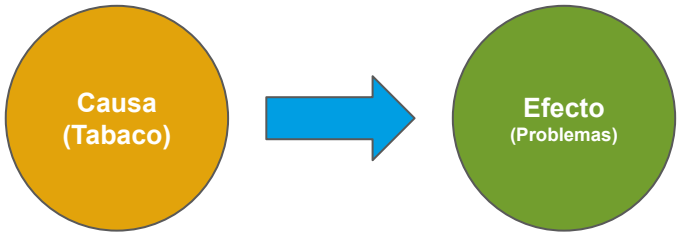
Diferencias

Causas y efectos son entidades ontológicamente diferentes.Las causas producen (o previenen) los efectos, pero los efectos no pueden producir las causas.

A esta asimetría lógica también le sigue una asimetría temporal: las causas siempre preceden en el tiempo a los efectos.



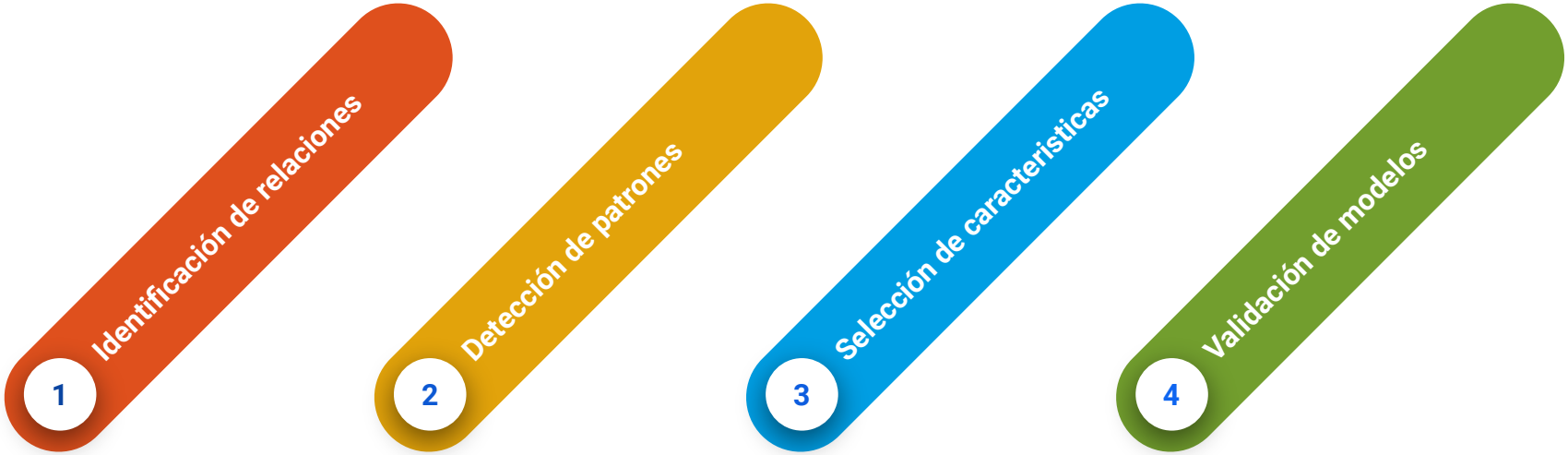
correlación



causalidad

Correlaciones

Análisis



/* Correlación lineal */

Correlación lineal

Covarianza

La Covarianza corresponde a una medida del grado de variación conjunta de dos o más variables respecto de sus medias.

Dados dos conjuntos de valores X e Y, su covarianza se calcula mediante la fórmula

$$Cov(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})}{n-1}$$

Correlación lineal

Coeficiente de correlación

La covarianza por sí sola puede ser difícil de interpretar porque su magnitud depende de las unidades en las que se miden las variables.

Podemos eliminar la influencia de las unidades dividiendo por las desviaciones estándar respectivas, y obtenemos el **coeficiente de correlación de Pearson (r)**

$$Cov(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{n-1}$$

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Correlación lineal

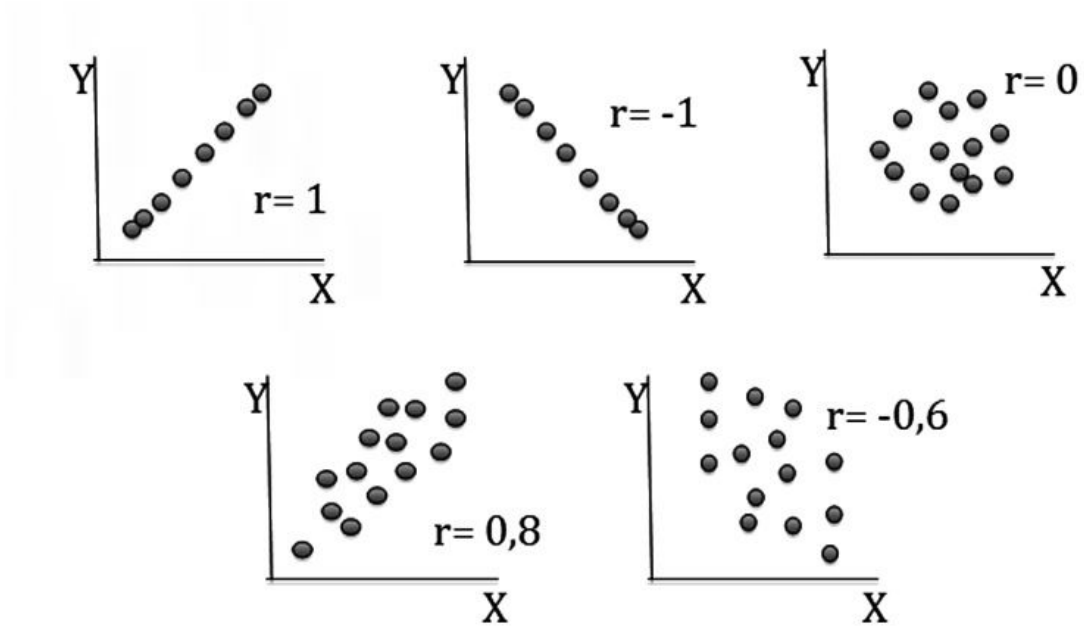
Correlación lineal y coeficiente

Si los puntos parecen seguir una forma de recta hablamos de **correlación lineal**, que es cuantificada mediante el **coeficiente de correlación**. Este toma valores entre -1 y 1, donde:

- **-1** indica una correlación negativa perfecta (una variable aumenta mientras la otra disminuye).
- **1** indica una correlación positiva perfecta (ambas variables aumentan o disminuyen juntas).
- **0** indica una correlación nula (no hay relación lineal entre las variables).

Coeficiente de correlación

Correlación lineal



¡Manos a la obra! - Matriz de correlaciones con Python

{desafío}
latam_

Matriz de correlaciones con Python

Correlaciones entre pares de variables

Veremos cómo analizar fácilmente las correlaciones entre pares de variables de un DataFrame utilizando Python. Para ello, puedes abrir tu propio archivo de Juýter Notebook y seguir los pasos que realizará tu profesor.

¡Manos al teclado!

{desafío} latam_

/*Regresión*/

Regresión

Concepto

La regresión es una técnica estadística utilizada para modelar la **relación** entre una variable dependiente y una o más variables independientes.

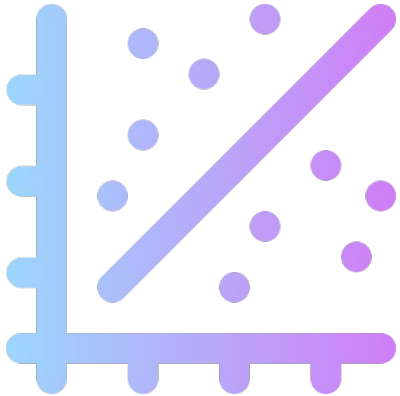
En otras palabras, se utiliza para **comprender cómo una variable cambia en respuesta a cambios en otras variables**.

Es una herramienta fundamental en el análisis de datos porque permite identificar y cuantificar relaciones entre variables, lo que a su vez puede ayudar a predecir valores futuros y tomar decisiones informadas.

Regresión

Regresión Lineal - dos variables

Su objetivo principal es encontrar **la mejor línea recta** que se ajusta a los datos, de manera que pueda utilizarse para predecir los valores de la variable dependiente en función de los valores de las variables independientes.



{desafío} latam_

Regresión

Regresión Lineal - dos variables

En la regresión lineal entre dos variables, la relación ellas variables se modela mediante la ecuación de una línea recta, que tiene la forma:

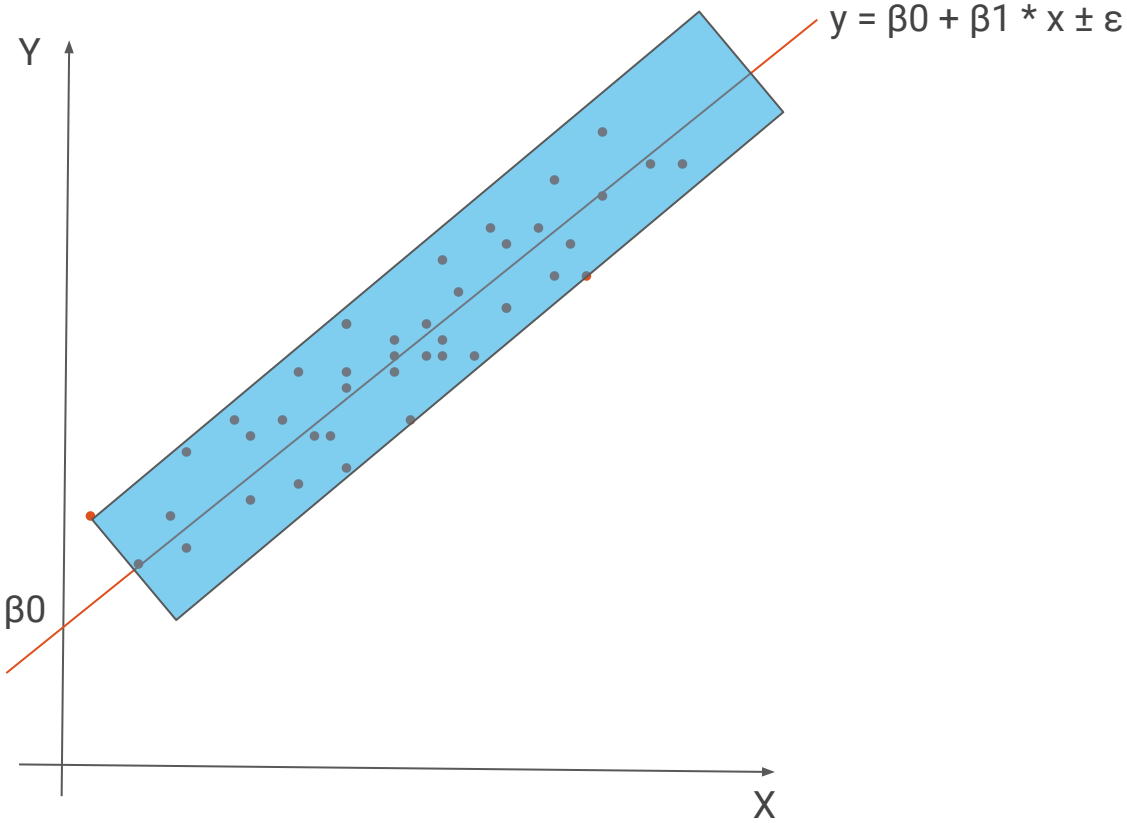
$$y = \beta 0 + \beta 1 * x \pm \varepsilon$$

Donde:

- y es la variable dependiente que estamos tratando de predecir.
- x es la variable independiente.
- β0 es la intersección de la recta con el eje Y, también llamada **intercepto**.
- β1 es la pendiente de la recta
- ε es el término de error, que tiene en cuenta las variaciones no explicadas por el modelo.

Regresión

Regresión Lineal



{desafío}
latam_

Regresión

Regresión lineal múltiple

Una extensión de la regresión lineal simple es la regresión múltiple, que permite modelar la relación entre una variable dependiente y dos o más variables independientes.

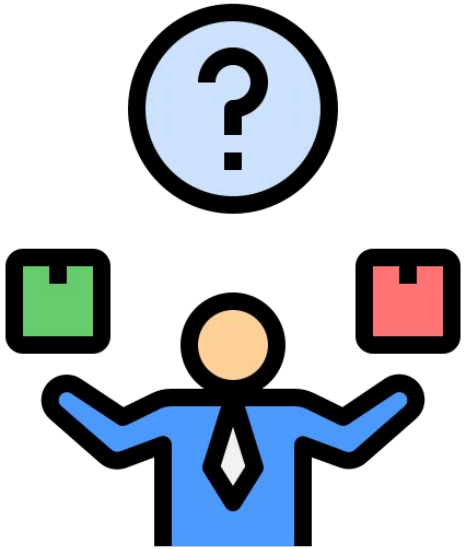
En lugar de considerar solo una variable independiente, como en la regresión lineal simple, la regresión múltiple tiene en cuenta múltiples factores que podrían influir en la variable dependiente. Se obtiene así un plano (o hiperplano) de la forma

$$y = \beta 0 + \beta 1 * x1 + \beta 2 * x2 + ... + \beta n * xn + \varepsilon$$

Regresión lineal

Residuos

Se llama **residuos** a las diferencias entre los valores observados y los valores predichos por el modelo de regresión. En otras palabras, son las discrepancias entre los puntos de datos reales y los puntos que se encuentran en la línea de regresión. Los residuos son importantes para evaluar la calidad del ajuste del modelo a los datos.



Regresión lineal

Mínimos cuadrados

Podemos calcular los cuadrados de los residuos y sumarlos (obteniendo un valor positivo o cero). Mientras menor sea este valor, significará que los cuadrados de los residuos son menores, y por ello que los residuos son menores.

¿Qué significa que los residuos sean menores? ¡Que el modelo es más preciso!

¡Manos a la obra!
Regresión lineal con Python

{desafío}
latam_

Regresión lineal con Python

Buscando la mejor recta

Vamos a ver cómo aplicar esto con la ayuda de Python. Para ello, puedes abrir tu propio archivo de Jupyter Notebook para replicar los pasos que te mostrará tu profesor.

/*Analizando el modelo*/

Análisis del Modelo

Resultados de la regresión

Naturalmente, necesitaremos evaluar lo adecuado o bien ajustado que sea nuestro modelo. Para esto contaremos con algunas métricas, que revisarán en la Tutoría.

Desafío
“Regresión Lineal”

{desafío}
latam_

Desafío

"Regresión lineal"

- Descarga el archivo "Desafío".
- Tiempo de desarrollo asincrónico: desde 2 horas.
- Tipo de desafío: individual.

¡AHORA TE TOCA A TI! 💪

{desafío} latam_

Ideas fuerza

*La **correlación** nos indica el grado de **interrrelación** entre dos o más variables, y nos permite suponer eventual **causalidad** entre ellas aunque no necesariamente será así.*

*La **regresión lineal** es un modelo que nos permite **relacionar variables** independientes con una dependiente, mediante una recta o un hiperplano*

*Podemos **evaluar un modelo** de regresión por medio de **métricas e indicadores,** que nos señalan cuánto se ajusta a los datos y su efectividad*

¿Qué aspectos consideras más relevante en la regresión lineal?

Recursos asincrónicos

¡No olvides revisarlos!

Para esta semana deberás revisar:

- Guía de estudio.
- Desafío “Regresión lineal”

Próxima sesión...

Generar modelos predictivos utilizando scikit-learn de acuerdo a requerimientos

{desafío} latam_

{desafío}
latam_

Academia de
talentos digitales