Correlación

Correlación lineal

La correlación es una medida estadística que nos permite cuantificar el grado de relación o dependencia entre dos variables. Para el caso de una dependencia lineal entre las variables, se obtiene con el coeficiente de correlación de Pearson:

$$\rho_{xy} = \frac{Cov_{xy}}{\sigma_x \sigma_y}$$

Donde:

Cov_{xy} es la covarianza entre los valores de las variables ***x*** e ***y***.

σ_x y σ_y son las desviaciones estándar de las variables ***x*** e ***y***.

Consideraciones: el cálculo del coeficiente de correlación de Pearson es para variables cuantitativas continuas que tienen una distribución normal y que presenten homocedasticidad. Es sensible a los valores atípicos.

Los valores que toma el coeficiente de Pearson están en el intervalo:

$$1 \leq \rho_{xy} \leq 1$$

En la siguiente tabla se puede ver el significado de la relación según el rango de valores.

Valor de ρ	Correlación
-1 ≤ ρ ≤ -0.8	Negativa fuerte.
-0.8 ≤ ρ ≤ -0.5	Negativa moderada
-0.5 ≤ ρ ≤ 0	Negativa débil
0	Sin correlación
0 ≤ ρ ≤ 0.5	Positiva débil
0.5 ≤ ρ ≤ 0.8	Positiva moderada
0.8 ≤ ρ ≤ 1	Positiva fuerte

Coeficiente de Spearman

Es el equivalente al coeficiente de Pearson cuando los datos son ordinales o cuando no cumplen la condición de normalidad necesaria. Además, se transforman los datos a rangos. Se calcula de la siguiente forma:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Donde:

d_i es la distancia entre los rangos de cada observación ($x_i - y_i$).

n es el número de observaciones.

La interpretación de los parámetros es la misma, pero se agrega el término de error.

Consideraciones: la relación entre las variables debe ser monótona.

Estimación con R

Para encontrar la correlación entre dos variables, vamos a utilizar las funciones de la instalación base de R: *cor()* y *cor.test()*.

Los parámetros que vamos a utilizar para las dos funciones son:

x: vector, matriz o data.frame numérico.

y: vector, matriz o data.frame numérico de la misma dimensión que x (opcional).

method: donde le indicamos que coeficiente de correlación obtendremos, por defecto el valor es “pearson”, pero se pueden utilizar “spearman” y “kendall”.

Ejercicios

Ejercicio 1 – Se tienen los datos de peso y altura de 15 individuos y se quiere ver si existe una relación lineal entre ellas y de que tipo es esa relación.

Peso [kg]	Altura [cm]
74	168
92	196
63	170
72	175
58	162
76	183
85	169
78	190
67	172
91	188
85	186
73	176
62	170
80	176
72	179

Analice los datos y justifique su repuesta.

Ejercicio 2 – Cargue el archivo de datos bank.csv utilizado en la guía anterior. Encuentre si existe una relación lineal entre las variables “age” y “balance”. Justifique la respuesta.

Ejercicio 3 – Cargue el conjunto de datos “mammals” del paquete **MASS** que contiene los pesos del cuerpo y el cerebro de un conjunto de mamíferos.

Se quiere conocer la relación entre las dos variables.

a. Explore los datos. ¿Puede utilizar el coeficiente de Pearson?

b. Si los supuestos se cumplen, encuentre el coeficiente de correlación. Si no, utilice la rho de Spearman. ¿Qué tipo de correlación existe?

c. Realice un escalamiento de los datos con el objetivo de “mejorar” la distribución de los datos y repita el análisis de correlación.

Ejercicio – 4. El conjunto de datos “Orange” del paquete **datasets** contiene los valores de edad y circunferencia de 5 árboles de naranjas.

a. Explore los datos.

b. Encuentre la correlación entre las variables. ¿Qué relación encuentra entre las variables?

Referencias

1. “Correlación lineal y Regresión lineal simple”. Disponible en: https://www.cienciadedatos.net/documentos/24_correlacion_y_regresion_lineal

2. “An Introduction to corrplot Package”. Disponible en: https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html