correlación\_lineal

2024-03-16

library(readxl)  
data <- as.data.frame(read\_excel("C:/Ejercicios/correlacion\_lineal/data.xls"))  
View(data)  
print(data)

## id peso altura  
## 1 1 23.47 34.72  
## 2 2 5.81 87.59  
## 3 3 67.92 16.28  
## 4 4 12.36 52.94  
## 5 5 38.09 11.67  
## 6 6 54.73 65.83  
## 7 7 89.24 28.15  
## 8 8 16.57 43.26  
## 9 9 73.10 78.94  
## 10 10 42.85 25.09  
## 11 11 6.28 59.37  
## 12 12 91.63 37.62  
## 13 13 29.04 20.48  
## 14 14 50.17 91.75  
## 15 15 18.92 6.82  
## 16 16 81.36 74.56  
## 17 17 37.59 48.29  
## 18 18 64.28 13.75  
## 19 19 10.73 82.36  
## 20 20 45.81 39.21  
## 21 21 79.05 57.83  
## 22 22 27.36 22.67  
## 23 23 58.29 83.94  
## 24 24 14.67 30.17  
## 25 25 33.92 63.45  
## 26 26 70.18 47.92  
## 27 27 8.73 18.26  
## 28 28 96.51 70.83  
## 29 29 21.84 9.51  
## 30 30 47.36 54.29

#Actividad 1:

La correlación lineal es una medida estadística que evalúa la relación lineal entre dos variables cuantitativas. Es una medida de la fuerza y la dirección de la relación entre estas dos variables. Cuando dos variables tienen una correlación lineal positiva, significa que a medida que una variable aumenta, la otra también tiende a aumentar. Por otro lado, una correlación lineal negativa indica que a medida que una variable aumenta, la otra tiende a disminuir.

La correlación lineal se representa mediante el coeficiente de correlación, comúnmente denotado como “r”. Este coeficiente oscila entre -1 y 1. Un valor de 1 indica una correlación lineal perfecta positiva, -1 indica una correlación lineal perfecta negativa, y 0 indica que no hay correlación lineal entre las dos variables.

Es importante tener en cuenta que la correlación lineal no implica causalidad. Es decir, el hecho de que dos variables estén correlacionadas linealmente no significa necesariamente que una cause la otra; puede haber otros factores involucrados que influyan en la relación entre las variables

#Actividad 2:

La correlación lineal se considera una prueba de correlación paramétrica porque asume que los datos siguen una distribución específica, típicamente una distribución normal, y que tienen una relación lineal entre sí. Las pruebas paramétricas, en general, requieren que los datos cumplan ciertas suposiciones sobre la distribución de la población subyacente y la naturaleza de los datos.

Las pruebas paramétricas son aquellas que se basan en parámetros específicos de la población subyacente, como la media y la varianza. Estas pruebas son más poderosas cuando se cumplen las suposiciones subyacentes y pueden proporcionar estimaciones más precisas de los parámetros poblacionales.

Por otro lado, las pruebas no paramétricas son aquellas que no hacen suposiciones sobre la distribución de la población subyacente. Estas pruebas son útiles cuando los datos no cumplen con los supuestos de normalidad u otras suposiciones paramétricas. Las pruebas no paramétricas son menos sensibles a las desviaciones de la normalidad y a otros supuestos, pero pueden ser menos potentes cuando se cumplen las suposiciones paramétricas.

En resumen, las pruebas paramétricas como la correlación lineal requieren ciertas suposiciones sobre la distribución de los datos y la naturaleza de la relación entre las variables, mientras que las pruebas no paramétricas son más flexibles en términos de suposiciones pero pueden ser menos potentes en ciertas situaciones.

#Actividad 3. Calcula la correlación entre las variables almacenadas en la tabla ‘data’.

correlacion\_datos <- cor(data)  
print(correlacion\_datos)

## id peso altura  
## id 1.000000000 0.08321171 0.003792534  
## peso 0.083211709 1.00000000 0.147768260  
## altura 0.003792534 0.14776826 1.000000000

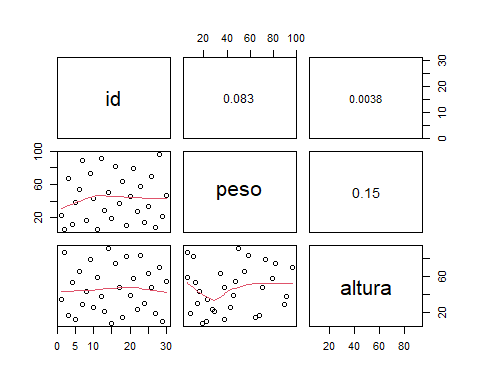
#Actividad 4. Calcula los coeficientes de correlación de las variables junto con el nivel de significancia (p-value) en 1 solo gráfico. Interpreta los resultados.

panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...) {  
 usr <- par("usr")  
 on.exit(par(usr))  
 par(usr = c(0, 1, 0 ,1))  
 Cor <- abs(cor(x, y))   
 txt <- paste0(prefix, format(c(Cor, 0.123456789), digits = digits)[1])  
 if(missing(cex.cor)) {  
 cex.cor <- 0.4 / strwidth(txt)  
 }  
 text(0.5, 0.5, txt,  
 cex = 1 + cex.cor\*Cor)  
}

#Aquí dibujamos la matriz de correlación

pairs(data,  
 upper.panel = panel.cor, # Este es el panel de correlación  
 lower.panel = panel.smooth)

## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter  
  
## Warning in par(usr): argument 1 does not name a graphical parameter



Actividad 5. ### Ejercicio 5. Emplea una función para obtener en una matriz de correlación lineal, IC 95% y p-value de todas las variables en el data frame ‘data’.

library(correlation)  
matriz <- correlation(data)  
matriz

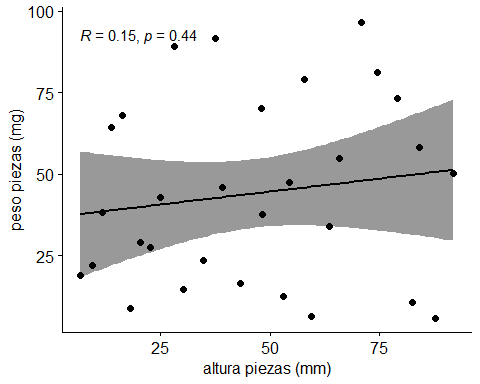
## # Correlation Matrix (pearson-method)  
##   
## Parameter1 | Parameter2 | r | 95% CI | t(28) | p  
## -------------------------------------------------------------------  
## id | peso | 0.08 | [-0.29, 0.43] | 0.44 | > .999  
## id | altura | 3.79e-03 | [-0.36, 0.36] | 0.02 | > .999  
## peso | altura | 0.15 | [-0.22, 0.48] | 0.79 | > .999  
##   
## p-value adjustment method: Holm (1979)  
## Observations: 30

### Ejercicio 6. Visualiza gráficamente la correlación lineal existente entre las variables ‘longitud’ y ‘peso’.

library(ggpubr)

## Loading required package: ggplot2

library(ggplot2)  
ggscatter(data, x = "altura", y = "peso",  
 add = "reg.line", conf.int = TRUE,  
 cor.coef = TRUE, cor.method = "pearson",  
 xlab = "altura piezas (mm)", ylab = "peso piezas (mg)")

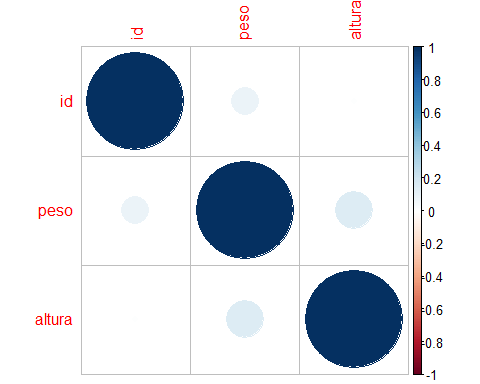


### Ejercicio 7. Emplea la librería ‘corrplot()’ para visualizar la correlación entre variables.

library(corrplot)

## corrplot 0.92 loaded

corrplot(cor(data))



### Ejercicio 8. A partir de la siguiente secuncia de valores numéricos:

#a) Creamos 2 vectores: ‘distancia’ y ‘n\_piezas’ para almacenarlos en un data frame:

distancia <- c( 1.1,100.2,90.3,5.4,57.5,6.6,34.7,65.8,57.9,86.1)  
n\_piezas <- c(110,2,6,98,40,94,31,5,8,10)  
datos\_2 <- data.frame(distancia, n\_piezas)  
print(datos\_2)

## distancia n\_piezas  
## 1 1.1 110  
## 2 100.2 2  
## 3 90.3 6  
## 4 5.4 98  
## 5 57.5 40  
## 6 6.6 94  
## 7 34.7 31  
## 8 65.8 5  
## 9 57.9 8  
## 10 86.1 10

#b) Calcula el coeficiente de correlación:

correlacion\_datos\_2 <- cor(datos\_2)  
print(correlacion\_datos\_2)

## distancia n\_piezas  
## distancia 1.0000000 -0.9249824  
## n\_piezas -0.9249824 1.0000000

#c) Calcula el nivel de significancia:

significancia\_datos\_2 <- cor.test(datos\_2$distancia, datos\_2$n\_piezas)$p.value  
print(significancia\_datos\_2)

## [1] 0.0001264706

#d) Calcula el Intervalo de confianza al 95% en relación con el coeficiente de correlación:

intervaloconfianza\_datos\_2 <- cor.test(datos\_2$distancia, datos\_2$n\_piezas)$conf.int  
print(intervaloconfianza\_datos\_2)

## [1] -0.9824414 -0.7072588  
## attr(,"conf.level")  
## [1] 0.95

#e) ¿Qué intensidad y dirección presentan ambas variables?

La intensidad es alta debido al acercarse a la cifra 1.

#f) ¿Es significativa esta relación?

Sí.

#g) Resulta apropiado afirmar la correlación (o no) entre variables con un tamaño muestral tan reducido (n=10).

No, ya que se necesita mas variabilidad de datos.

#9 Explícame con un ejemplo en R la diferencia entre una relación lineal y monótona entre 2 variables Primero, generaremos datos para ambos tipos de relaciones y luego visualizaremos los resultados utilizando gráficos de dispersión y ajustes de línea.

# Cargamos la librería necesaria

library(ggplot2)

# Creamos datos para una relación lineal

set.seed(123) # Para reproducibilidad x\_lineal <- 1:100 y\_lineal <- 2 \* x\_lineal + rnorm(100, mean = 0, sd = 10)

# Creamos datos para una relación monótona no lineal

x\_monotono <- 1:100 y\_monotono <- x\_monotono^2 + rnorm(100, mean = 0, sd = 300)

# Crear un dataframe para cada tipo de relación

datos\_lineales <- data.frame(x = x\_lineal, y = y\_lineal) datos\_monotonos <- data.frame(x = x\_monotono, y = y\_monotono)

# Gráfico de dispersión y ajuste lineal para relación lineal

ggplot(datos\_lineales, aes(x, y)) + geom\_point() + geom\_smooth(method = “lm”, se = FALSE, color = “blue”) + # Ajuste lineal labs(title = “Relación Lineal”, x = “Variable X”, y = “Variable Y”)

# Gráfico de dispersión y ajuste lineal para relación monótona

ggplot(datos\_monotonos, aes(x, y)) + geom\_point() + geom\_smooth(method = “lm”, se = FALSE, color = “red”) + # Ajuste lineal labs(title = “Relación Monótona”, x = “Variable X”, y = “Variable Y”)

En este ejemplo, generamos dos conjuntos de datos. Uno para una relación lineal y otro para una relación monótona no lineal. Luego, creamos gráficos de dispersión y ajustes de línea utilizando ggplot2. La diferencia clave entre las dos relaciones radica en cómo cambian las variables dependientes (Y) con respecto a las variables independientes (X).

En la relación lineal, los puntos se distribuyen alrededor de una línea recta. El ajuste lineal (línea azul) muestra la tendencia lineal entre las variables. A medida que X aumenta, Y aumenta o disminuye de manera constante.

En la relación monótona, los puntos siguen una tendencia en forma de curva (en este caso, una parábola). El ajuste lineal (línea roja) no captura adecuadamente la relación entre las variables, ya que no es lineal. En este caso, a medida que X aumenta, Y puede aumentar o disminuir, pero sigue una dirección general sin cambios abruptos en la dirección de la relación.

Estos gráficos ilustran claramente la diferencia entre una relación lineal y una relación monótona en un conjunto de datos.

# 10) ¿Qué tipo de prueba de correlación se aplica a las variables que experimentan una relación monótona? Expón un ejemplo en R.

Cuando las variables muestran una relación monótona, es más apropiado utilizar pruebas de correlación no paramétricas. Una de las pruebas más comunes para evaluar la correlación en este contexto es la prueba de correlación de Spearman. Esta prueba calcula el coeficiente de correlación de Spearman, que mide la relación monotónica entre dos variables.

# Ejemplo de prueba de correlación de Spearman en R

# Cargamos la librería necesaria

library(stats)

# Creamos datos para una relación monótona

set.seed(123) # Para reproducibilidad x\_monotono <- 1:100 y\_monotono <- x\_monotono^2 + rnorm(100, mean = 0, sd = 300)

# Creamos un dataframe con los datos

datos\_monotonos <- data.frame(x = x\_monotono, y = y\_monotono)

# Realizamos la prueba de correlación de Spearman

cor\_spearman <- cor.test(datos\_monotonosy, method = “spearman”)

# Mostramos los resultados

print(cor\_spearman)

En este ejemplo, generamos datos para una relación monótona no lineal y luego aplicamos la prueba de correlación de Spearman utilizando la función cor.test() con el parámetro method = “spearman”. Esta función calcula el coeficiente de correlación de Spearman y realiza una prueba de hipótesis para determinar si la correlación es significativa.

El resultado de la prueba de correlación de Spearman incluirá el valor del coeficiente de correlación de Spearman (rho), así como el valor p asociado, que indica la significancia estadística de la correlación encontrada.

Es importante recordar que la prueba de correlación de Spearman es adecuada para evaluar la correlación entre variables cuando la relación no es necesariamente lineal, pero sigue una tendencia monotónica.