Actividad Correlación lineal

Manuel Martín

2024-03-11

library(readxl)  
data <- as.data.frame(read\_excel("C:/Users/manueltristan/Documents/actividadcorrelacionlineal/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.

Hablamos de una medida estadística empleada para analizar la relación existente entre dos variables de caracter cuantitativo. Sobre esta relación, la correlación lineal permite conocer su intensidad, desde 0 (no existe correlación) a 1 (correlación total) y también la dirección de esa correlación. Cuando es un número positivo hablamos de correlación directamente proprocional y si el número es negativo hablamos de correlación inversamente proporcional.

### Actividad 2

Hablamos de una correlación paramétrica porque se asume una distribución de Gauss o distribución normal y que la relación entre las variables analizadas es lineal. Unas asunciones sobre la distribución de los datos y naturaleza de la relación, que no se tienen que producir en las pruebas no paramétricas.

### Actividad 3

Calculamos la correlación entre las variables de la tabla de datos.

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

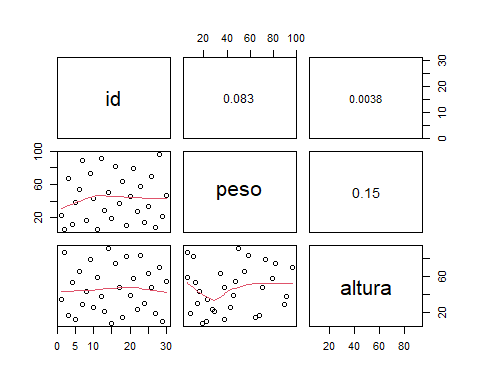
## id peso altura  
## id 1.000000000 0.08321171 0.003792534  
## peso 0.083211709 1.00000000 0.147768260  
## altura 0.003792534 0.14776826 1.000000000

### Actividad 4

Calculamos los coeficientes de correlación de las variables y el nivel de significancia en un mismo gráfico.

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)  
}  
#Dibujamos la matriz de correlación  
pairs(data,  
 upper.panel = panel.cor, # 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



En el gráfico se observan las correlaciones existentes entre las variables. Mostrando que no existe una correlación entre las variables analizadas, al menos no una correlación lineal.

### Ejercicio 5

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

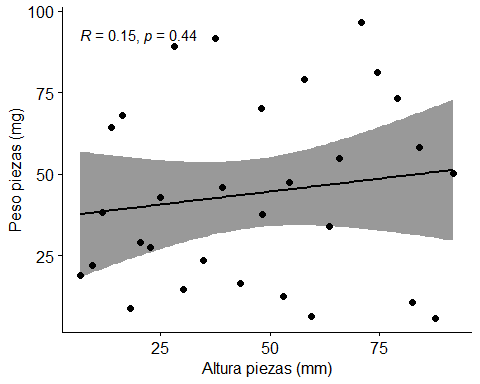
Los resultados obtenidos muestran un coeficiente de correlación entre las variables igual o inferior a 0.15. También en todos los casos el p-value o nivel de significancia es superior a 0.05, lo que significa que no tenemos suficiente evidencia para rechazar que la correlación sea resultado del azar y no sea significativa. A esto debemos incluir que el intervalo de confianza en los tres casos incluye el 0. Por lo que podemos afirmar que no hay una correlación significativa entre las variables en los datos analizados.

### Ejercicio 6

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)")

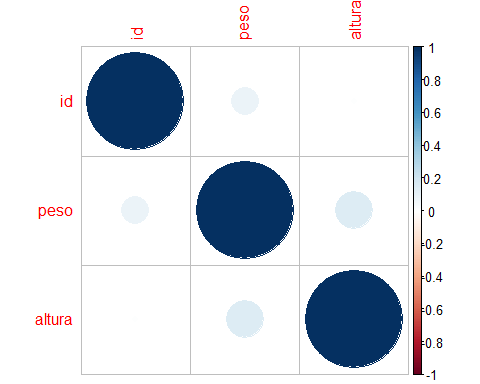


### Ejercicio 7

library(corrplot)

## corrplot 0.92 loaded

corrplot(cor(data))



### Ejercicio 8

1. Creamos los dos vectores para 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

1. Calculamos 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

1. Calculamos 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

1. Calculamos el Intervalo de confianza al 95%:

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

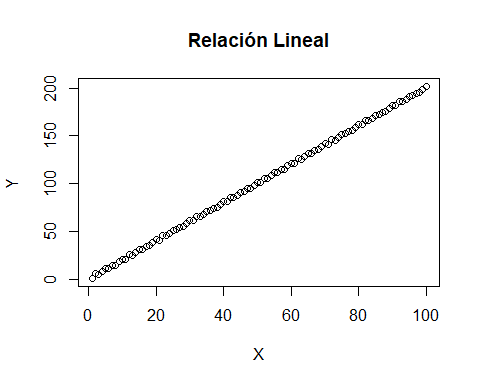
1. Ambas variables presentan una intensidad de 0.9249824, cercana a 1 y una dirección negativa, que significa que la correlación es inversa, cuando aumenta una de las variables la otra disminuye.
2. La relación es significativa estadisticamente hablando porque el p\_value o nivel de significancia es inferior al 0.05. Los 0.0001264706 obtenidos justifican pensar que la correlación obtenida no es resultado del azar y que es significativa.
3. Aunque el nivel de significancia permita afirmar que estamos ante una relación significativa, trabajar con 10 muestras suele ser escaso y por lo general se suele recomendar un tamaño muestral mayor (30-100) sería más apropiado.

### Ejercicio 9

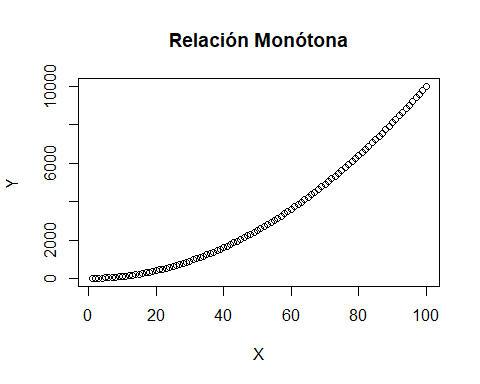
La diferencia entre una relación lineal y una monótona es que la relación lineal se puede representar mediante una línea recta en un gráfico de dispersión mientras que la relación monótona no tiene porque ser representada mediante una línea recta, sino que la dirección de la relación sea constante. Lo vemos con un ejemplo:

##### Ejemplo de relación lineal

x <- 1:100  
y <- 2\*x + rnorm(10)  
plot(x, y, main = "Relación Lineal", xlab = "X", ylab = "Y")

 ##### Ejemplo de relación monótona

x <- 1:100  
y <- x^2 + rnorm(10)  
plot(x, y, main = "Relación Monótona", xlab = "X", ylab = "Y")



### Ejercicio 10

Para las variables con una relación monótona podemos emplear la correlación de Spearman o la prueba de correlación de Kendall.

##### Ejemplo de prueba de correlación de Spearman

x <- 1:100  
y <- x^2 + rnorm(10)  
correlation\_spearman <- cor.test(x, y, method = "spearman")  
print(correlation\_spearman)

##   
## Spearman's rank correlation rho  
##   
## data: x and y  
## S = 0, p-value < 2.2e-16  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 1