TAREA 2

Se nos ha asignado una base de datos que contiene 120 observaciones de 14 variables, la cual se nos ha pedido responder con la ayuda de R a una serie de cuestiones que resolveremos a continuación.

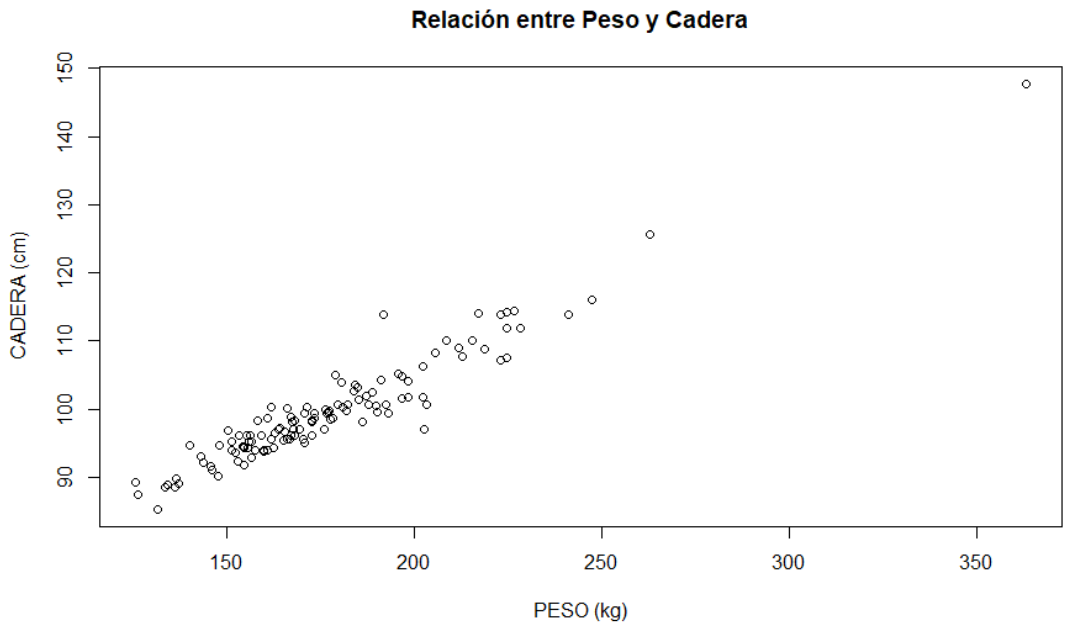
Antes de nada, primero hemos de cargar las librerías que emplearemos a lo largo del ejercicio y también importar la base de datos asignada:

```
<R>
library(haven)
library(MASS)
library(fitdistrplus)
library(survival)
library(dplyr)
datos <- read_sav("_Pallarés_Díez.sav")
```

1. Elige dos variables, X e Y, que presenten una correlación superior a 0.79.

En este paso, se calculó la matriz de correlación para identificar las variables con una correlación mayor a 0.79. Las dos variables seleccionadas fueron 'Peso' (X) y 'Cadera' (Y), con una correlación de 0.956.

Al hacer una representación de ambas variables obtuvimos el siguiente resultado:



```
<R>
# Calcular la matriz de correlación
correlaciones <- cor(datos, use = "complete.obs")
# Filtrar correlaciones mayores a 0.79
correladas <- which(correlaciones > 0.79, arr.ind = TRUE)

# Crear un dataframe para ver las variables correlacionadas
resultado_correlaciones <- data.frame(
 Variable_X = names(datos)[correladas[, 1]],
 Variable_Y = names(datos)[correladas[, 2]],
 Correlacion = correlaciones[correladas]
)
# Eliminar pares donde X e Y son la misma variable
resultado_correlaciones <- resultado_correlaciones %>%
 filter(Variable_X != Variable_Y)

# Mostrar las correlaciones filtradas
print(resultado_correlaciones)

# Elegir un par de variables
peso<-datos$Peso
cadera<-datos$Cadera
plot(peso,cadera)
```

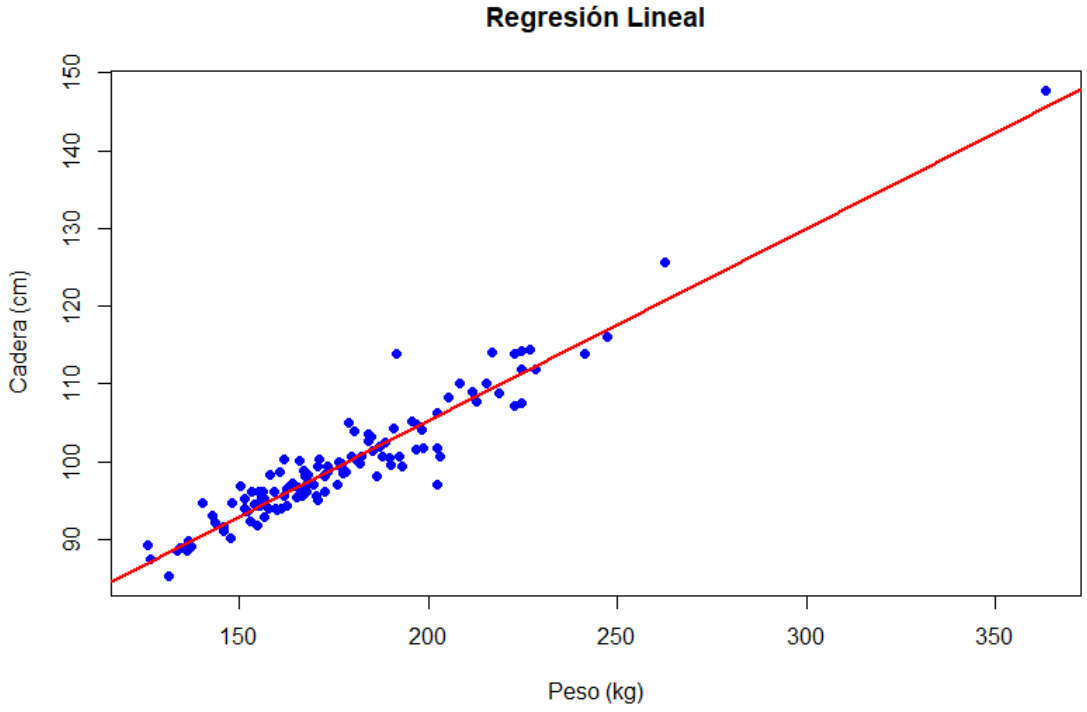
2. Estima el modelo lineal Y = β0 + β1 · X. Considera su "residual standard error".

```
<R>
# Ajustar el modelo lineal entre las variables seleccionadas
modelo <- lm(cadera ~ peso)

# Mostrar el resumen del modelo
resumen_modelo <- summary(modelo)
print(resumen_modelo)

# Coeficientes estimados
beta_0 <- resumen_modelo$coefficients[1, 1]  # Intercepto
beta_1 <- resumen_modelo$coefficients[2, 1]  # Pendiente

# Error estándar residual
sigma <- resumen_modelo$sigma
#Mu= β^0 + β^1 · X
mu<-beta_0 + beta_1*peso

# Graficar los datos y la línea de regresión
# Graficar los datos y la línea de regresión
plot(peso, cadera, main = "Regresión Lineal",
    xlab = "Peso (kg)", ylab = "Cadera (cm)", pch = 19, col = "blue")
abline(modelo, col = "red", lwd = 2)
```

```
Call:
lm(formula = cadera ~ peso)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9294 -1.3474 -0.1321  1.2115 10.5258

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.91250    1.25560   44.53   <2e-16 ***
peso         0.24700    0.00697   35.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.41 on 118 degrees of freedom
Multiple R-squared:  0.9141,    Adjusted R-squared:  0.9134
F-statistic:  1256 on 1 and 118 DF,  p-value: < 2.2e-16
```

β0 β1 σ

El modelo lineal ajustado es: Y = β0 + β1 * X, donde se utilizaron las variables 'Peso' y 'Cadera'. El resultado muestra que el intercepto es 55.91 y la pendiente es 0.247, con un error estándar residual de 2.41.



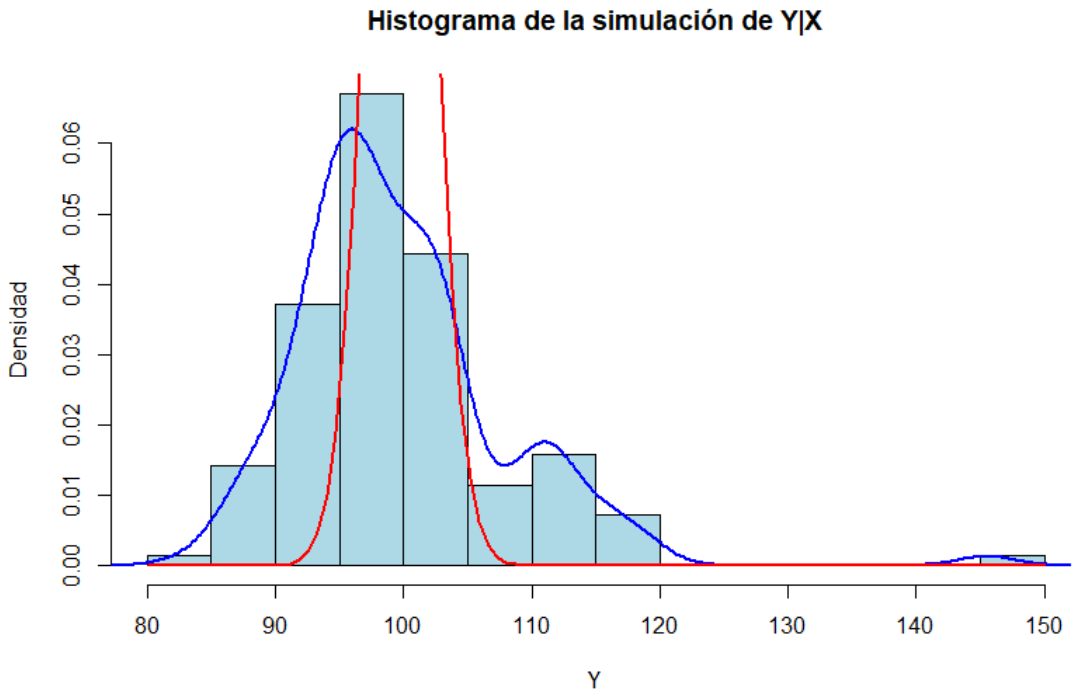
La imagen muestra un gráfico de dispersión de la relación entre "Peso" (en kg) y "Cadera" (en cm), junto con una línea de regresión lineal ajustada en color rojo. Cada punto azul representa un individuo con su correspondiente valor de "Peso" y "Cadera", y se observa una tendencia positiva, lo que indica que a medida que el peso aumenta, el tamaño de la cadera también tiende a aumentar. La línea roja representa el ajuste del modelo lineal Y = β0 + β1 * X, donde Y es "Cadera" y X es "Peso", mostrando una pendiente positiva que confirma la relación positiva entre ambas variables. El gráfico sugiere que existe una fuerte correlación lineal entre "Peso" y "Cadera", lo cual es consistente con el coeficiente de correlación elevado obtenido previamente.

3. Considera la variable Y |X. Comprueba, con una muestra simulada, si se puede aceptar que Y |X ∼ N (β^ 0 + β^ 1 · X, σ^)

Se simuló una muestra de Y dado X, utilizando los parámetros obtenidos del modelo. Se generó un histograma para visualizar la distribución simulada y se comparó con la normalidad esperada.

```
<R>
#Muestra simulada de Y |X ∼ N(β^0 + β^1 · X, σ^)
Y_sim <- rnorm(140, mean = mu, sd = sigma)


# Graficar el histograma de la simulación
hist(Y_sim, probability = TRUE, main = "Histograma de la simulación de Y|X",
    xlab = "Y", ylab = "Densidad", col = "lightblue", border = "black")


# Añadir la curva de densidad estimada
z <- density(Y_sim)
lines(z, col = "blue", lwd = 2)


# Añadir la línea de densidad teórica
curve(dnorm(x, mean = mean(Y_sim), sd = sigma), add = TRUE, col = "red", lwd = 2)


#Comprobación mediante Fitdis (AIC y BIC) si la normal es un modelo que se ajusta a los datos
fgamma <- fitdist(Y_sim, "gamma")
fexp <- fitdist(Y_sim, "exp")
fnorm <- fitdist(Y_sim, "norm")
fchi <- fitdist(Y_sim, "chisq", start = list(df = 2))
funif <- fitdist(Y_sim, "unif")
fweibull <- fitdist(Y_sim, "weibull")
flognorm <- fitdist(Y_sim, "lnorm")
```

Con la muestra simulada primero graficamos el histograma de la simulación añadiéndole la curva de la densidad estimada y la línea de densidad teórica para ver si se ajustaría a un modelo normal.



Aquí en la imagen vemos en color rojo la línea de densidad teórica y en azul la curva de densidad estimada, vemos que más o menos coinciden, pero no del todo por lo que no podemos aceptar ni rechazar la hipótesis de que siga una distribución normal.

Para ello con la librería fitdistrplus calculamos para diferentes modelos los valores BIC y AIC de nuestra muestra simulada para poder aceptar o rechazar la hipótesis de:

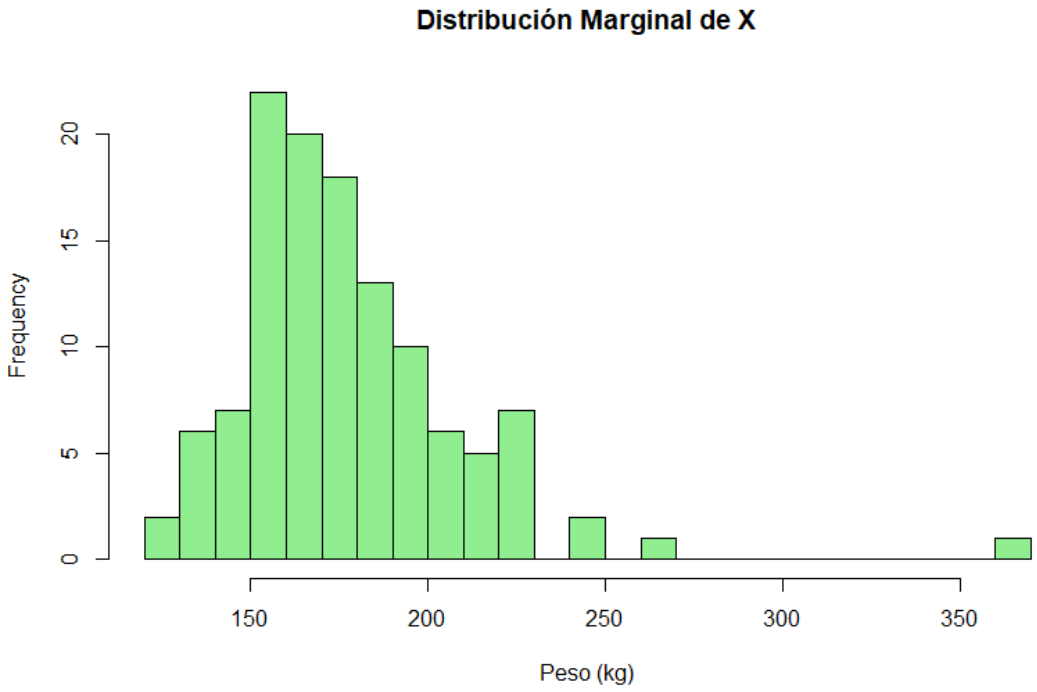
$$Y|X \sim N(\beta\hat{} 0 + \beta\hat{} 1 \cdot X, \sigma\hat{})$$

Comparamos los valores AIC y BIC para los modelos Gamma, Exponencial, Normal, Log-Normal, Uniforme, Chi cuadrado y Weibull y obtuvimos unos valores parecidos tanto como para Gamma, Normal y Log-normal. Por lo que no podemos rechazar entonces que nuestros datos sigan una distribución normal.

Modelo	AIC	BIC
Gamma	972.4538	978.3371
Exponential	1570.0640	1573.0050
Normal	981.9527	987.8360
Chi-Squared	1041.4580	1044.4000
Uniform	1147.5140	1153.3970
Weibull	1050.6170	1056.5000

4. Calcula la distribución marginal de X.

```
<R>

# Histograma de la variable X para visualizar la distribución marginal

hist(peso, breaks = 30, main = "Distribución Marginal de X", xlab = "peso", col = "lightgreen")
```



El histograma muestra la distribución marginal de la variable X (Peso). Se observa que la mayoría de los valores de peso están concentrados entre 140 y 220 kg, con una frecuencia máxima alrededor de los 150-170 kg. La distribución muestra una forma asimétrica con una ligera cola hacia la derecha, lo que sugiere un sesgo positivo. Esto indica que hay más individuos con pesos menores y algunos pocos con pesos significativamente mayores.

Para justificar si la distribución pudiese ser normal, se debe considerar la forma del histograma. Aunque la distribución tiene una concentración central y disminuye en ambos extremos (lo cual es una característica de la normalidad), el sesgo hacia la derecha indica que no es perfectamente simétrica, lo que sugiere que puede no seguir exactamente una distribución normal.

5. A partir de esta distribución marginal, calcula E(X).

```
<R>

esperanza_x <- mean(peso)
```

Para justificar matemáticamente el cálculo de la esperanza (o valor esperado) de la variable X (Peso), se utiliza la definición de valor esperado para una variable aleatoria continua. La fórmula general es:

$$E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$$

donde fX(x)es la función de densidad de probabilidad de X. Sin embargo, dado que trabajamos con datos discretos, podemos aproximar el valor esperado utilizando la media aritmética de los datos. La fórmula para la esperanza en este contexto es:

$$E(X) \approx \frac{1}{n} \sum_{i=1}^{n} x_i$$

donde:

- n es el número total de observaciones,
- xi representa cada valor de la variable X (Peso).

En el caso de los datos proporcionados, se ha calculado la media aritmética de todos los valores de "Peso" en la muestra, lo que equivale a sumar todos los valores de la variable y dividir entre el número total de observaciones. Esto da como resultado E(X)≈177.36, lo cual es una estimación del valor promedio del "Peso" en la muestra.

Matemáticamente, esto equivale a la aproximación del valor esperado para distribuciones de datos reales, donde la integral de la densidad se reemplaza por la suma discreta de los valores observados.

6. Señala la función de densidad de probabilidad conjunta de X e Y.

La función de densidad de probabilidad conjunta de X e Y, asumiendo que Y|X~N(β^0+β^1·X, σ^), es:

$$f(x, y) = f_X(x) \cdot f_{Y|X}(y|x)$$

Donde:

- f X(x) es la densidad marginal de X (que asumimos como una distribución empírica basada en los datos del "Peso").
- F Y|X (y | x) es la densidad condicional de Y dado X, que sigue una distribución normal con media β^0+β^1·x y desviación estándar σ^.

La forma explícita para la densidad condicional f Y|X (y | x) es:

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{(y - (\hat{\beta}_0 + \hat{\beta}_1 \cdot x))^2}{2\hat{\sigma}^2} \right)$$

La función de densidad conjunta completa sería:

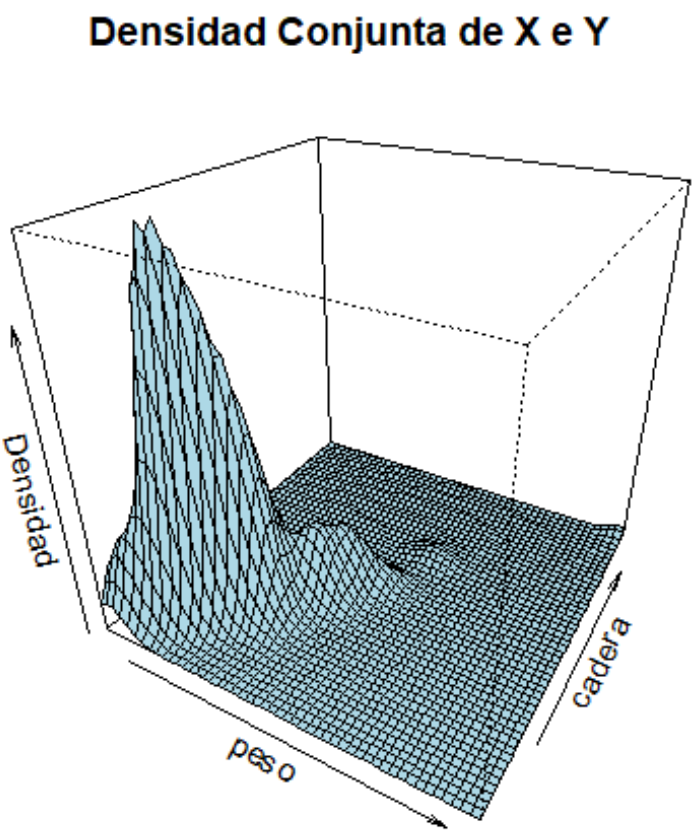
$$f(x,y) = f_X(x) \cdot \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{(y-(\hat{\beta}_0+\hat{\beta}_1 \cdot x))^2}{2\hat{\sigma}^2}\right)$$

Donde:

- β ^0=55.91
- β^1=0.247
- σ^=2.40

Esta expresión describe la probabilidad conjunta de los valores de "Peso" y "Cadera" bajo los supuestos del modelo ajustado.

```
<R>

# Calcular los parámetros necesarios para la función de densidad conjunta

mu_X <- mean(peso)

mu_Y <- mean(cadera)

sigma_X <- sd(peso)

sigma_Y <- sd(cadera)

rho <- cor(peso, cadera)


# Mostrar la función de densidad conjunta con los parámetros calculados

cat("La función de densidad conjunta de X e Y es:\n")

cat("f(x, y) =", 1 / (2 * pi * sigma_X * sigma_Y * sqrt(1 - rho^2)), "* exp(-1 / (2 * (1 -", rho^2, ")) * [((x -", mu_X, ") /", sigma_X, ")^2 + ((y -", mu_Y, ") /", sigma_Y, ")^2 - 2 *", rho, "* ((x -", mu_X, ") /", sigma_X, ") * ((y -", mu_Y, ") /", sigma_Y, "))])\n")


# Para estimar la densidad conjunta, usaremos la función kde2d del paquete MASS

densidad_conjunta <- kde2d(peso, cadera, n = 50)


# Graficar la densidad conjunta

persp(densidad_conjunta, main = "Densidad Conjunta de X e Y", xlab = "peso", ylab = "cadera", zlab = "Densidad", theta = 30, phi = 30, col = "lightblue")
```



Buscando encontré esta función en R para poder graficar la densidad conjunta gracias al paquete MASS.

La imagen muestra una gráfica tridimensional que representa la función de densidad conjunta de las variables "peso" y "cadera". En la gráfica, el eje horizontal representa los valores de "peso", el eje lateral representa los valores de "cadera" y el eje vertical muestra la densidad de probabilidad conjunta.

La superficie en la gráfica indica cómo varía la densidad de probabilidad conjunta en función de los valores de "peso" y "cadera". Se observa una mayor densidad en una región específica, lo que sugiere que los valores más comunes de "peso" y "cadera" se encuentran en esa zona alta. A medida que se aleja de esa región, la densidad disminuye, lo que indica que combinaciones de "peso" y "cadera" más alejadas de esta región central son menos frecuentes.

La forma de la superficie refleja cómo se distribuyen conjuntamente las dos variables, mostrando un pico en el área donde los valores de "peso" y "cadera" son más comunes, lo cual indica la mayor concentración de datos.