Actividad 7: Regresion

Jorge Eduardo de León Reyna - A00829759

```
# Importando dataset

data = read.csv("/content/Estatura-peso_HyM.csv")
head(data)
```

A data.frame: 6 × 3

	Estatura	Peso	Sexo
	<dbl>	<dbl>	<chr>
1	1.61	72.21	H
2	1.61	65.71	H
3	1.70	75.08	H
4	1.65	68.55	H
5	1.72	70.77	H
6	1.63	77.18	H

```
# Modificando valores de variable categorica
M = data
M$Sexo <- as.numeric(M$Sexo == "H")
```

```
head(M)
```

A data.frame: 6 × 3

	Estatura	Peso	Sexo
	<dbl>	<dbl>	<dbl>
1	1.61	72.21	1
2	1.61	65.71	1
3	1.70	75.08	1
4	1.65	68.55	1
5	1.72	70.77	1
6	1.63	77.18	1

La recta de mejor ajuste

1. Obtén la matriz de correlación de los datos que se te proporcionan. Interpreta.

```
correlation_matrix <- cor(M)
print(correlation_matrix)
```

```
            Estatura      Peso      Sexo
Estatura 1.0000000 0.8032449 0.5835090
Peso     0.8032449 1.0000000 0.7708846
Sexo     0.5835090 0.7708846 1.0000000
```

En esta matriz de correlación podemos ver que las variables que tienen un coheficiente de correlacion muy fuerte (>0.8) segun las aproximaciones de Cohen son "Peso" y "Estatura". Con correlacion fuerte (>0.5, <0.8) segun esta misma clasificación son las variables "Sexo" y "Peso" y "Sexo" y "Estatura".

2. Obtén medidas (media, desviación estándar, etc) que te ayuden a analizar los datos.

```
MM = subset(M, M$Sexo == "0")
MH = subset(M, M$Sexo == "1")
M1 = data.frame(MH$Estatura, MH$Peso, MM$Estatura, MM$Peso)
```

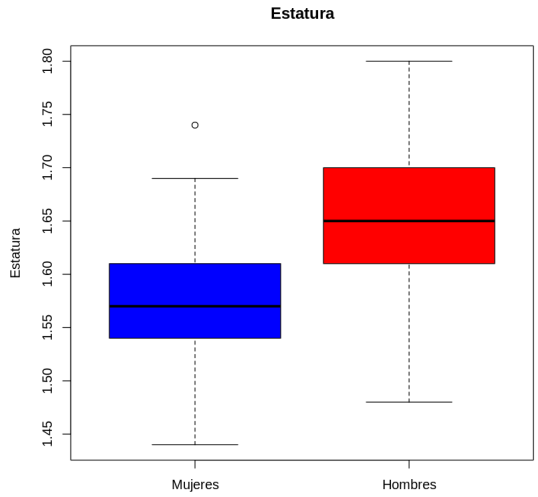
```
n = 4
d = matrix(NA,ncol=7,nrow=n)
for(i in 1:n){
  d[i,] = c(as.numeric(summary(M1[,i])), sd(M1[,i]))
}

m = as.data.frame(d)
```

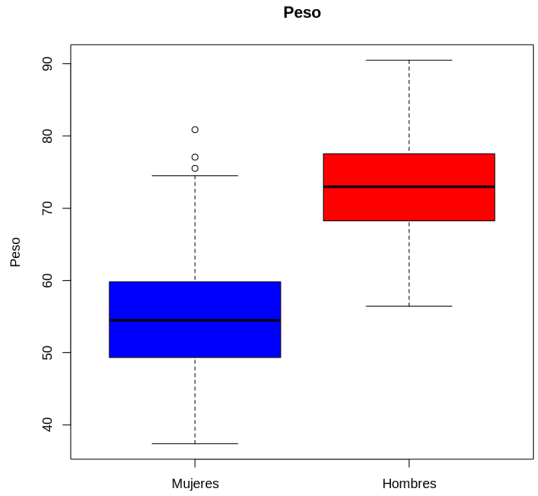
```
row.names(m) = c("H-Estatura", "H-Peso","M-Estatura", "M-Peso")
names(m) = c("Minimo", "Q1", "Mediana", "Media", "Q3","Maximo", "Deviacion Estandar")
m
```

A data.frame: 4 × 7

	Minimo	Q1	Mediana	Media	Q3	Maximo	Deviacion Estandar
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
H-Estatura	1.48	1.6100	1.650	1.653727	1.7000	1.80	0.06173088
H-Peso	56.43	68.2575	72.975	72.857682	77.5225	90.49	6.90035408
M-Estatura	1.44	1.5400	1.570	1.572955	1.6100	1.74	0.05036758
M-Peso	37.39	49.3550	54.485	55.083409	59.7950	80.87	7.79278074

```
#boxplot con la distribucion de los datos

boxplot(M$Estatura~M$Sexo, ylab="Estatura", xlab="", col =c("blue","red"), names = c("Mujeres", "Hombres"), main = "Estatura")
boxplot(M$Peso~M$Sexo, ylab="Peso", xlab="", col =c("blue","red"), names = c("Mujeres", "Hombres"), main = "Peso")
```





3. Escribe la ecuación de regresión de mejor ajuste.

3.1 Regresion del modelo

```
A = lm(M$Peso~M$Estatura+M$Sexo)
A
```

```
Call:
lm(formula = M$Peso ~ M$Estatura + M$Sexo)

Coefficients:
(Intercept)   M$Estatura       M$Sexo
     -85.32        89.26        10.56
```

```
b0 = A$coefficients[1]
b1 = A$coefficients[2]
b2 = A$coefficients[3]

cat("Peso = ", b0, "+ (", b1, " * Estatura) + (", b2, " * Sexo)|")
```

```
Peso =  -85.31907 + ( 89.26035  * Estatura) + ( 10.56447  * Sexo)|
```

3.2 Verificación del modelo

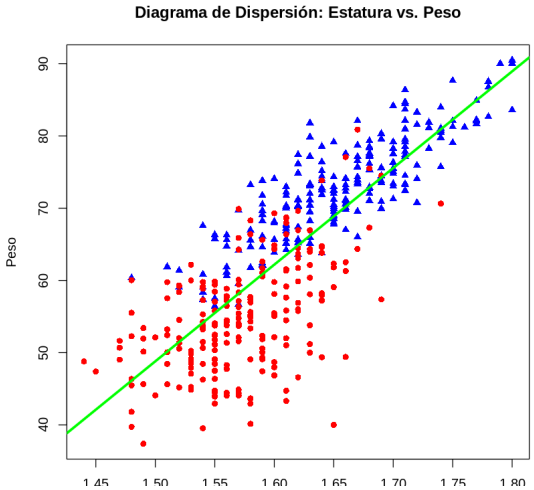
```
summary(A)
```

```
Call:
lm(formula = M$Peso ~ M$Estatura + M$Sexo)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9505  -3.2491   0.0489   3.2880  17.1243

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -85.3191     7.1874  -11.87   <2e-16 ***
M$Estatura   89.2604     4.5635   19.56   <2e-16 ***
M$Sexo       10.5645     0.6317   16.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.381 on 437 degrees of freedom
Multiple R-squared:  0.7837,    Adjusted R-squared:  0.7827
F-statistic: 791.5 on 2 and 437 DF,  p-value: < 2.2e-16
```

1. El modelo es significativo dado que se obtuvo un valor P = 2.2e-16 el cual es menor a alfa (0.03).
2. B1 es significativa dado que se obtuvo un valor P = 2.2e-16 el cual es menor a alfa (0.03).
3. El porcentaje de variacion explicada por el modelo es el 78.27% acorde al valor de R Cuadrada Ajustada.

4. Dibuja el diagrama de dispersión y la recta de mejor ajuste.

```
# Crear un diagrama de dispersión con diferentes colores para cada sexo
plot(M$Estatura, M$Peso,
     main = "Diagrama de Dispersión: Estatura vs. Peso",
     xlab = "Estatura",
     ylab = "Peso",
     pch = ifelse(M$Sexo == 0, 16, 17),
     col = ifelse(M$Sexo == 1, "blue", "red"),
     )

# Llínea de regresión ajustada
abline(lm(Peso ~ Estatura, data = data), col = "green", lwd=3)
```



5. Interpreta en el contexto del problema:

1. ¿Qué información proporciona β̂0 sobre la relación entre la estatura y el peso de hombres y mujeres?

- B0 es la intersección en nuestro modelo por lo que se puede interpretar como que cuando la estarura es 0, el peso sera este valor, esto no tiene mucho sentido en el contexto actual por lo que su uso en este caso solo es para calculos dentro del modelo.

2. ¿Cómo interpretas β̂1 en la relación entre la estatura y el peso de hombres y mujeres?

- B1 en este caso nos indica una proporcion directa entre la estatura y el peso teniendo como factor de proporcion el valor de B1.

Validacion del modelo

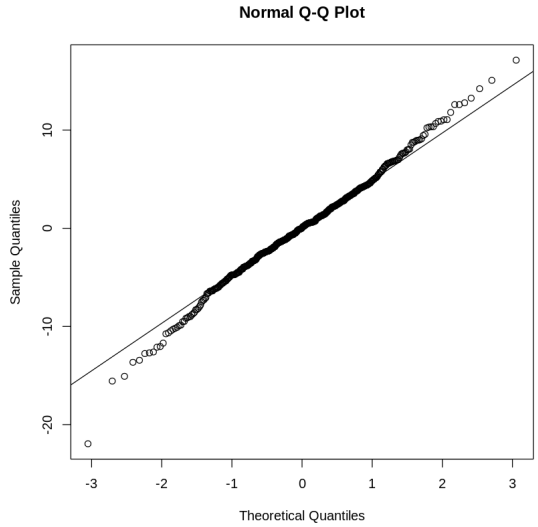
2. Analiza si el (los) modelo(s) obtenidos son apropiados para el conjunto de datos.

- Realiza el análisis de los residuos:
  - Normalidad de los residuos
  - Verificación de media cero
  - Homocedasticidad e independencia

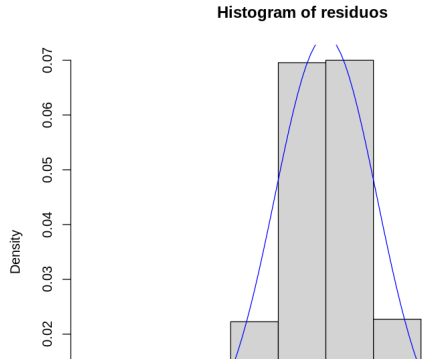
2.1 Normalidad de los residuos

1. Hipotesis: H0 = normalidad en la distribucion de errores, H1 = no normalidad en la distribucion de errores
2. Regla de decision: Si el valor P > 0.05 no se rechaza la hipotesis incicial por lo que se comprueba su normalidad

```
# Normalidad de los residuos

#1. Hipotesis: H0 = normalidad en la distribucion de errores, H1 = no normalidad en la distribucion de errores (P value > 0.05

residuos <- resid(A)

# Q-Q Plot
qqnorm(residuos)
qqline(residuos)

# Histograma
hist(residuos, probability = TRUE)
curve(dnorm(x, mean = mean(residuos), sd = sd(residuos)), add = TRUE, col = "blue")

# Prueba de Kolmogorov-Smirnov
ks.test(residuos, "pnorm", mean(residuos), sd(residuos))
```




```
Warning message in ks.test.default(residuos, "pnorm", mean(residuos), sd(residuos)):
“ties should not be present for the Kolmogorov-Smirnov test”

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  residuos
D = 0.03675, p-value = 0.5922
alternative hypothesis: two-sided
```



3. Analisis del resultado: despues de hacer la prueba Kolmogorov-Smirnov y complementarlo con la representacion grafica de las distribuciones podemos encontrar un valor P = 0.5922 el cual es mayor a 0.05.
4. Conclusión: debido que valor P > 0.05 no se rechaza la H0 y se afirma que los datos se distribuyen normalmente.

2.2 Verificacion de media 0

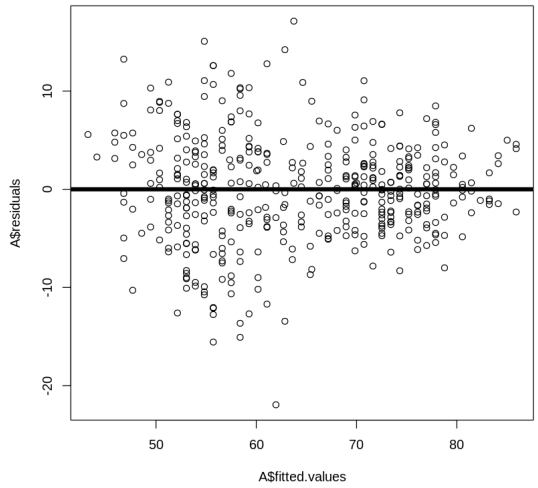
```
1 t.test(A$residuals)
```

```
        One Sample t-test

data:  A$residuals
t = 7.7026e-17, df = 439, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.5029859  0.5029859
sample estimates:
   mean of x
1.971277e-17
```

Con esta prueba podemos confirmar una media muy cercana a 0 al tener un valor P grande lo cual hace que no se rechaze la hipotesis inicial de tener una media de 0 o muy cercana a 0.

2.3 Homocedasticidad

```
plot(A$fitted.values,A$residuals)
abline(h=0, lwd=5)
```



En el grafico anterioer se puede ver una distribucion uniforme de los datos al rededor de la recta trazada lo cual sugiere que si se presenta Homocedasticidad en los residuos del modelo. Esto se interpreta como que los residuos estan normalmente distribuidos al rededor de una media 0 (la linea central).

Conclusion general e interpretación

En este modelo se pudo describir de manera satisfactoria el comportamiento de los datos al hacer pruebas que verificaran su validez. Estas pruebas tanto visuales como a travez de tests estadisticos corroboraron la validez del modelo en diferentes aspectos, tales como la distribucion de sus errores, la presesencia de Homocedasticidad y las pruebas de significancia tanto para los coeficientes individuales del modelo como para este mismo a nivel general.

Con esto se puede afirmar entonces que se logro modelas satisfactoriamente la relacion entre la estatura y el sexo para el calculo de el peso en una persona.