Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

M1. Regresión Lineal

TC3006C.101 Inteligencia artificial avanzada para la ciencia de datos I

Profesores:

Ivan Mauricio Amaya Contreras

Blanca Rosa Ruiz Hernandez

Antonio Carlos Bento

Frumencio Olivas Alvarez

Hugo Terashima Marín

Alumno:

Alberto H Orozco Ramos - A00831719

31 de Agosto de 2023

Instrucciones

Analiza la base de datos de estatura y peso de los hombres y mujeres en México y obten el mejor modelo de regresión para esos datos.

Verificación del Modelo con Estaturas de Hombres y Mujeres

Para regresión lineal de estatura (m) y peso (kg) de hombres y mujeres mexicanos, analiza si el modelo es útil para el conjunto de datos:

- Significación del modelo
- Significación individual (de $\beta 1$)
- Coeficiente de determinación

Importamos Google Drive

In [ ]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

In [ ]:
```
# Cargamos el lenguaje de R para utilizarlo en Google Colab
%load_ext rpy2.ipython
```

Cargamos las librerías necesarias

In [ ]:
```
%%R
install.packages('nortest')

library(nortest)
```

Cargamos los datos desde el archivo CSV

In [ ]:
```
%%R
M = read.csv("/content/drive/MyDrive/Colab Notebooks/Estadística/Estatura-peso_HyM.
```

1. La recta de mejor ajuste (Primera entrega)

1. Obtén la matriz de correlación de los datos que se te proporcionan. Interpreta.
2. Obtén medidas (media, desviación estándar, etc) que te ayuden a analizar los datos.
3. Encuentra la ecuación de regresión de mejor ajuste:
4. Realiza la regresión entre las variables involucradas

5. Verifica el modelo: 1. Verifica la significancia del modelo con un alfa de 0.03. 2. Verifica la significancia de β̂i con un alfa de 0.03. 3. Verifica el porcentaje de variación explicada por el modelo
6. Dibuja el diagrama de dispersión de los datos y la recta de mejor ajuste.
7. Interpreta en el contexto del problema cada uno de los análisis que hiciste.
8. Interpreta en el contexto del problema:
9. ¿Qué información proporciona β̂0 sobre la relación entre la estatura y el peso de hombres y mujeres?
10. ¿Cómo interpretas β̂1 en la relación entre la estatura y el peso de hombres y mujeres?

Medidas para analizar datos

Esto incluye valor mínimo, Q1, mediana, media, Q3, máximo y desviación estándar

In [ ]:
```
%%R

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

n=4 #número de variables
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","Máximo","Desv Est")
m
```

```
           Minimo      Q1 Mediana     Media      Q3 Máximo   Desv Est
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
```

Matriz de Correlación

Aqui podemos observar cómo algunas variables parecen ser más compatibles con unas que otras. Por ejemplo, la estatura de los hombres presentan una mayor correlación con el peso de hombres que con la estatura. El de las mujeres es igual, sin embargo la correlación EstaturaM-PesoM es menor que la de los hombres EstaturaH-PesoH:

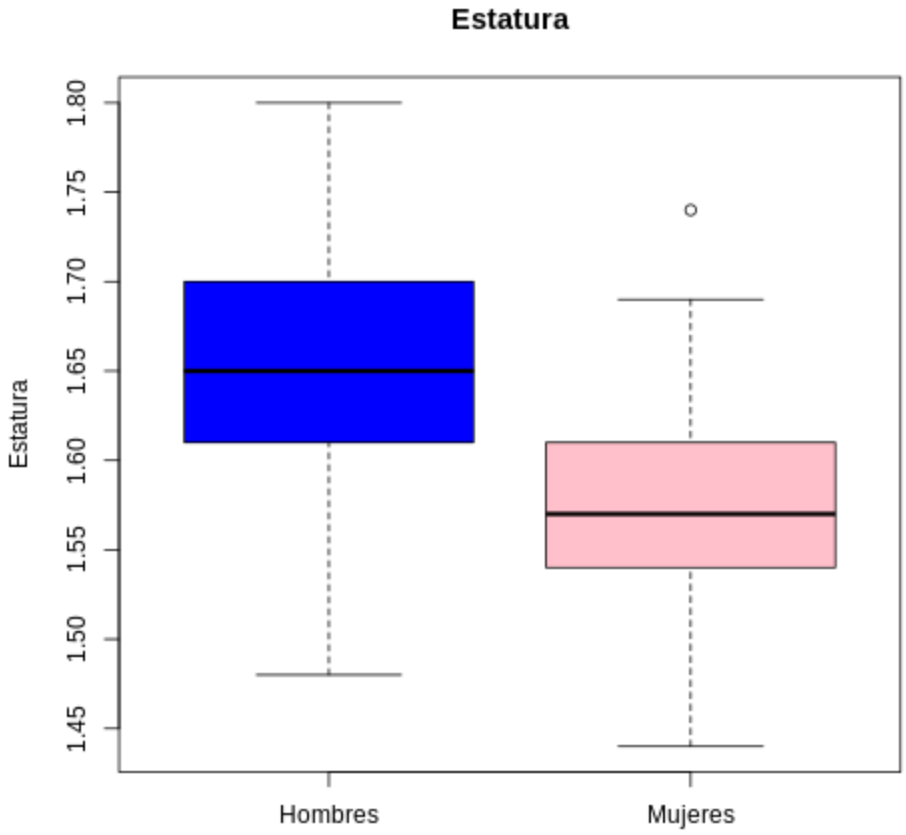
In [ ]:
```
%%R
print(cor(M1))
cat('\n')
cor(M$Estatura, M$Peso)
```

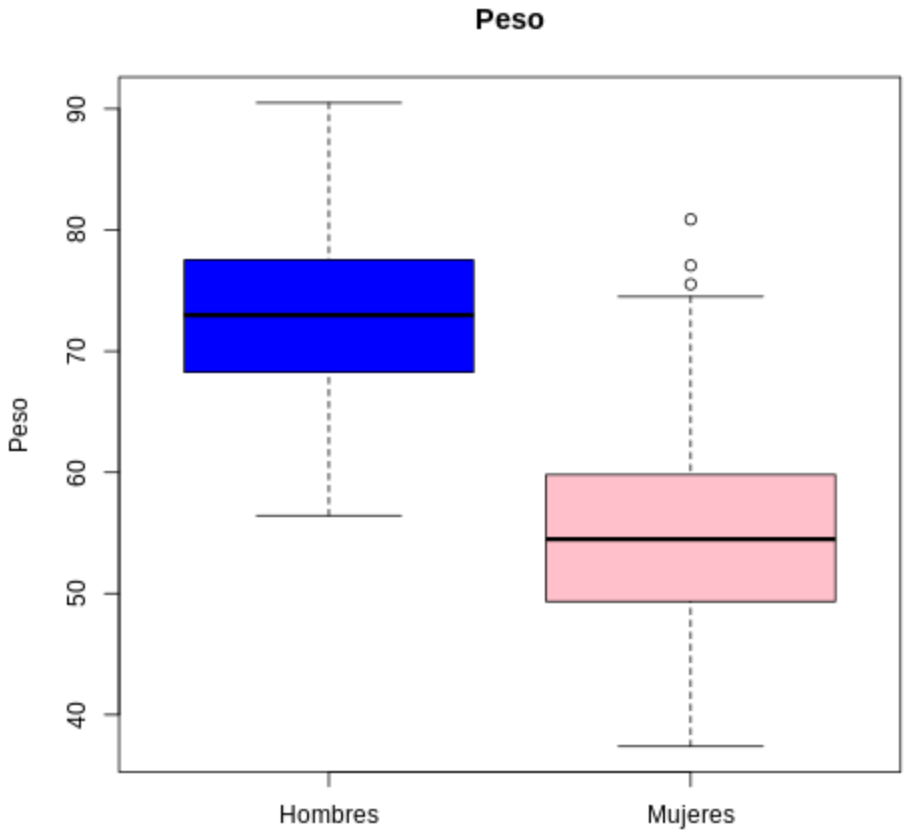
```
             MH.Estatura     MH.Peso  MM.Estatura    MM.Peso
MH.Estatura 1.0000000000 0.846834792 0.0005540612 0.04724872
MH.Peso     0.8468347920 1.000000000 0.0035132246 0.02154907
MM.Estatura 0.0005540612 0.003513225 1.0000000000 0.52449621
MM.Peso     0.0472487231 0.021549075 0.5244962115 1.00000000

[1] 0.8032449
```

Descripción de las variables mediante boxplots

In [ ]:
```
%%R
boxplot(M$Estatura~M$Sexo, ylab="Estatura", xlab="", col=c("blue","pink"), names=c(
boxplot(M$Peso~M$Sexo, ylab="Peso",xlab="", names=c("Hombres", "Mujeres"), col=c("b
```





Regresión lineal entre las variables involucradas

In [ ]:
```
%%R
A = lm(M$Peso~M$Estatura+M$Sexo)
A
```

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

Coefficients:
(Intercept)   M$Estatura      M$SexoM
     -74.75        89.26       -10.56
```

Extraemos los coeficientes de la regresión lineal realizada

In [ ]:
```
%%R
b0 = A$coefficients[1]
b1 = A$coefficients[2]
b2 = A$coefficients[3]

cat("Peso =", b0,"+", b1, "Estatura", b2, "SexoM")
```

```
Peso = -74.7546 + 89.26035 Estatura -10.56447 SexoM
```

Verificación del modelo

- Significancia global
- Significación individual
- Porcentaje de variación explicada por el modelo

In [ ]:
```
%%R
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) -74.7546     7.5555  -9.894   <2e-16 ***
M$Estatura   89.2604     4.5635  19.560   <2e-16 ***
M$SexoM     -10.5645     0.6317 -16.724   <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
```

Ecuaciones del Modelo

Para Mujeres (Sexo=1)

In [ ]:
```
%%R
cat("Para Mujeres", "\n")
cat("Peso =", b0 + b2, "+", b1, "Estatura", "\n")

cat("Para hombres", "\n")
cat("Peso =", b0, "+", b1, "Estatura", "\n")
```

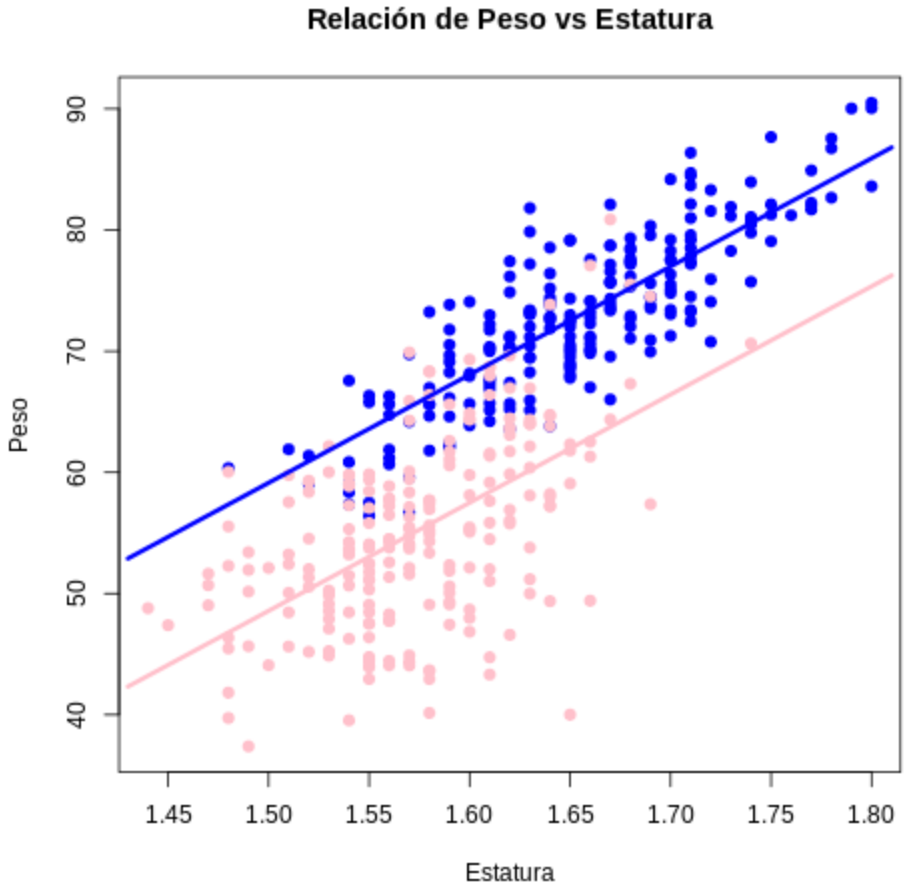
```
Para Mujeres
Peso = -85.31907 + 89.26035 Estatura
Para hombres
Peso = -74.7546 + 89.26035 Estatura
```

Diagrama de dispersión de los datos y rectas de mejor ajuste

In [ ]:
```
%%R
Ym = function(x){b0+b2+b1*x}
Yh = function(x){b0+b1*x}

colores = c("blue", "pink")
```

```
plot(M$Estatura, M$Peso, col=colores[factor(M$Sexo)], pch=19, ylab = "Peso", xlab =
x = seq(1.43, 1.81, 0.01)
lines(x, Ym(x), col="pink", lwd=3)
lines(x, Yh(x), col="blue", lwd=3)
```



Interpretación dentro del contexto del problema:

1. ***¿Qué información proporciona β0 sobre la relación entre la estatura y el peso de hombres y mujeres?***
   La intercepción de la relación estatura-peso de los hombres resulta más alta que el de las mujeres a pesar de tratarse de 2 pendientes muy similares y paralelas entre sí. El hecho de que una pendiente se encuentra una encima de la otra indica que los hombres en promedio tienden a tener pesos más altos que mujeres con su misma altura. Básicamente eso es lo que β0 nos puede decir de estas relaciones.
2. ***¿Cómo interpretas β1 en la relación entre la estatura y el peso de hombres y mujeres?***
   Considero que tomando en cuenta a b1, el coeficiente de estatura representa el cambio de peso asociado con un cambio de unidad de altura. Es decir, cada vez que aumenta una unidad de altura, el cambio esperado en peso sería de b1 unidades. Otro dato bastante curioso es que b1 tiene el mismo valor de 89.26035 para ambos casos, hombres y mujeres.

Modelo con interacción

In [ ]:
```
%%R
B = lm(M$Peso~M$Estatura*M$Sexo)
B
```

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

Coefficients:
       (Intercept)          M$Estatura             M$SexoM  M$Estatura:M$SexoM
            -83.68               94.66               11.12               -13.51
```

In [ ]:
```
%%R
summary(B)
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-21.3256  -3.1107   0.0204   3.2691  17.9114

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -83.685      9.735  -8.597   <2e-16 ***
M$Estatura           94.660      5.882  16.092   <2e-16 ***
M$SexoM              11.124     14.950   0.744    0.457
M$Estatura:M$SexoM  -13.511      9.305  -1.452    0.147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.374 on 436 degrees of freedom
Multiple R-squared:  0.7847,    Adjusted R-squared:  0.7832
F-statistic: 529.7 on 3 and 436 DF,  p-value: < 2.2e-16
```

2. Validación del Modelo (segunda entrega)

1. Retoma el notebook en el que realizaste el análisis de regresión que encontraste 'La recta de mejor ajuste'
2. Analiza si el (los) modelo(s) obtenidos son apropiados para el conjunto de datos. Realiza el análisis de los residuos:
3. Normalidad de los residuos
4. Verificación de media cero
5. Homocedasticidad e independencia
6. No te olvides de incluir los cuatro pasos en las pruebas de hipótesis que realices: (1) Hipótesis, (2) Regla de decisión, (3) Análisis del resultado, (4) Conclusión.
7. Interpreta en el contexto del problema cada uno de los análisis que hiciste.

8. Emite una conclusión final sobre el análisis de regresión lineal que conjunte lo que hiciste en las dos partes de esta actividad.

Prueba de Normalidad Anderson-Darling en los residuos del primer modelo

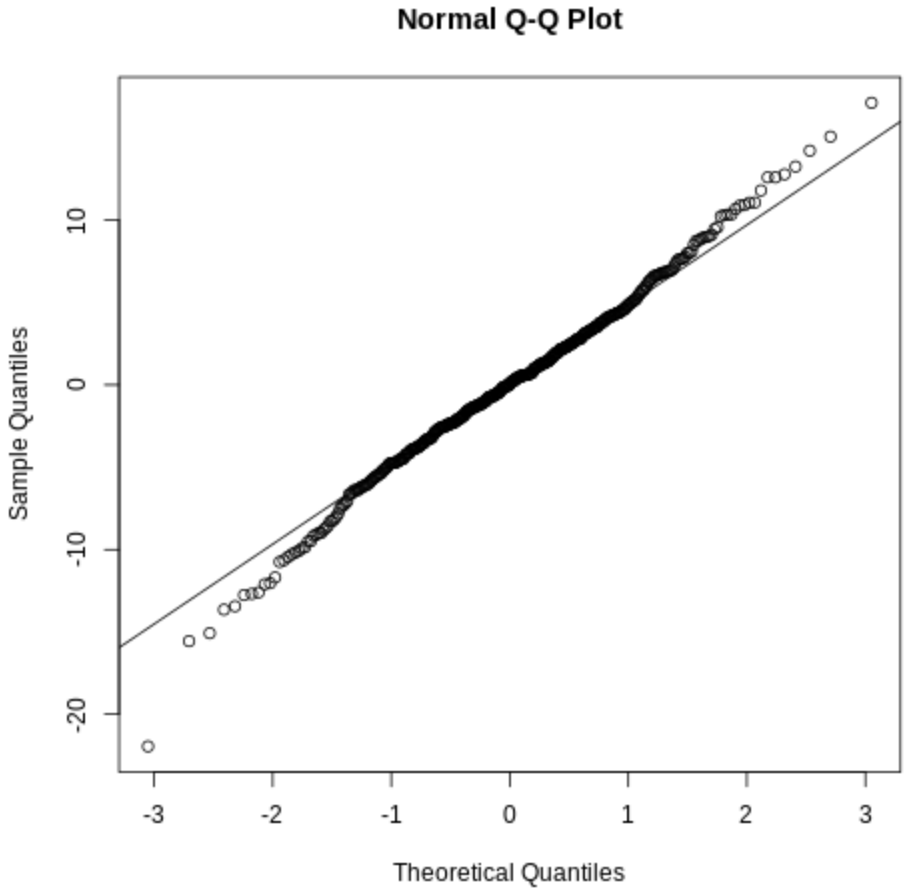
In [ ]:
```
%%R
ad.test(A$residuals)
```

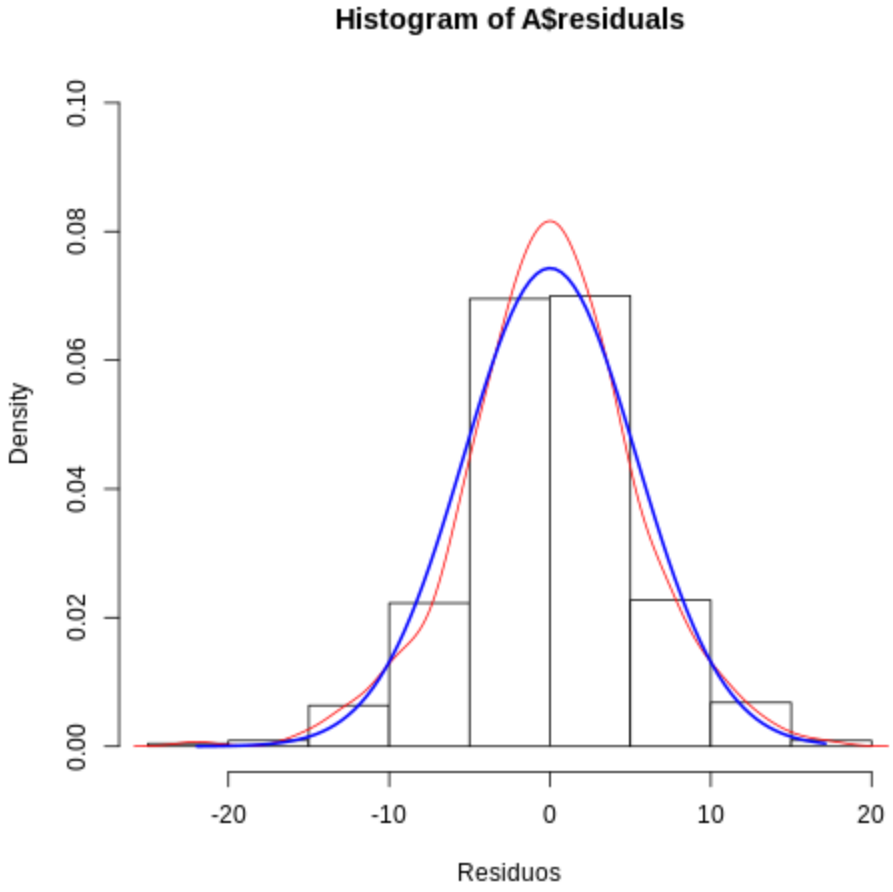
```
	Anderson-Darling normality test

data:  A$residuals
A = 0.79651, p-value = 0.03879
```

Gráfica QQPlot e Histograma de los residuos de los datos para normalidad

In [ ]:
```
%%R
qqnorm(A$residuals)
qqline(A$residuals)
hist(A$residuals,freq=FALSE, ylim = c(0, 0.1), xlab="Residuos", col=0)
lines(density(A$residuals),col="red")
curve(dnorm(x,mean=mean(A$residuals),sd=sd(A$residuals)), from=min(A$residuals),
to=max(A$residuals), add=TRUE, col="blue",lwd=2)
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
10
0
-10
-20
-3
-2
-1
0
1
2
3



Media Cero

La media verdadera según la prueba t de student dice que la media no es igual a 0, sin embargo el valor es muy aproximado.

In [ ]:
```
%%R
t.test(A$residuals)
```

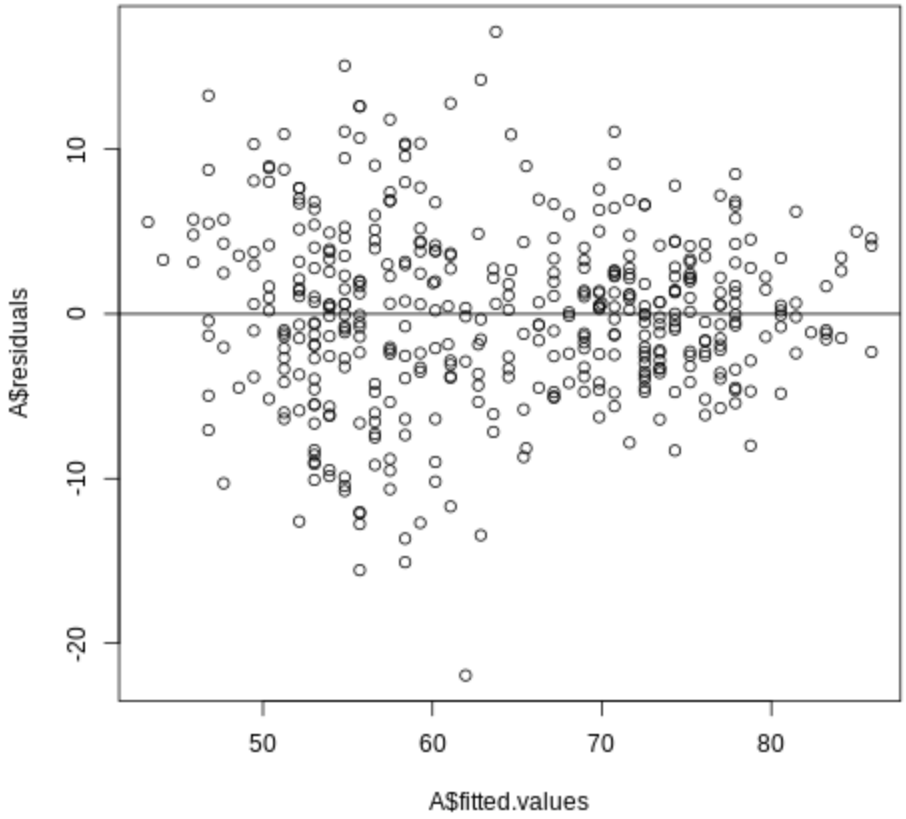
```
	One Sample t-test

data:  A$residuals
t = -3.3793e-16, 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
-8.648385e-17
```

Homocedasticidad

In [ ]:
```
%%R
plot(A$fitted.values, A$residuals)
```

```
abline(h=0, color="blue")
```

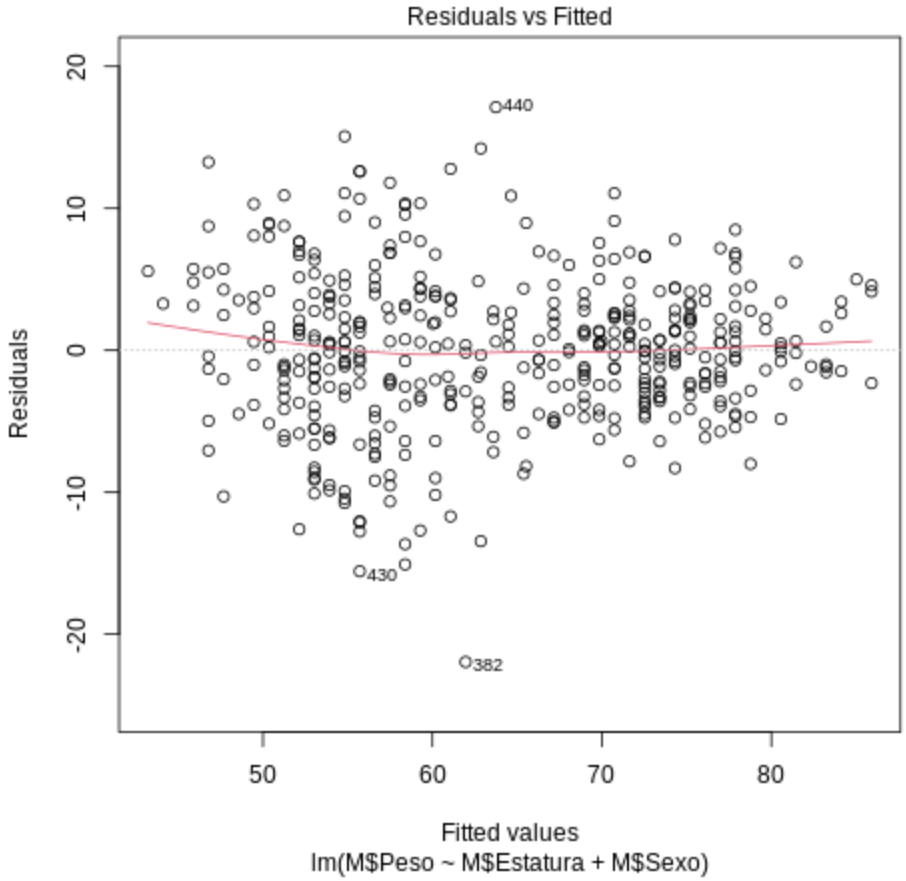


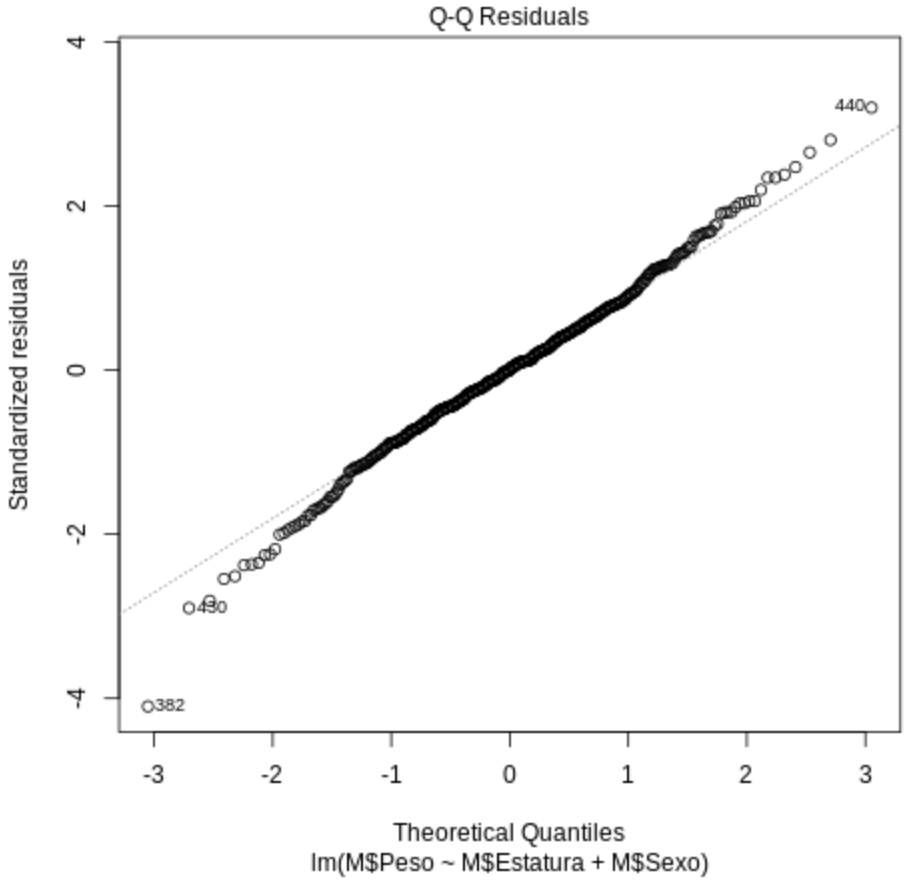
2. Intervalos de confianza (última entrega)

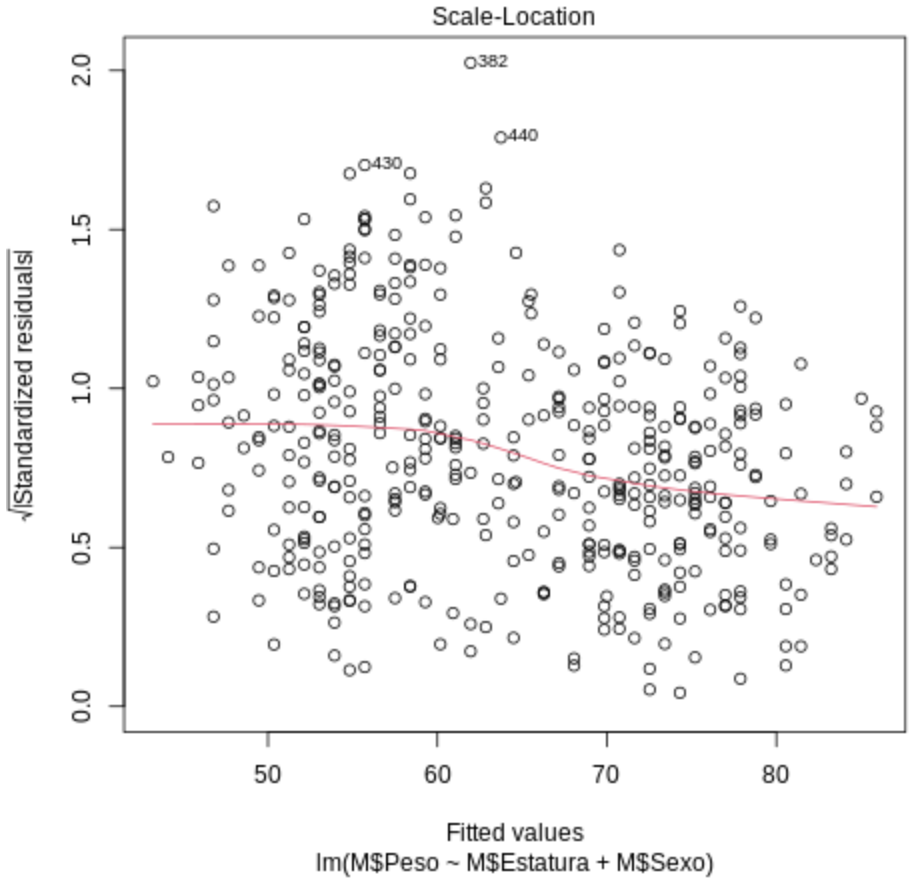
1. Con los datos de las estaturas y pesos de los hombres y las mujeres construye la gráfica de los intervalos de confianza y predicción para la estimación y predicción de Y para el modelo obtenido. Interpreta y comenta los resultados obtenidos.

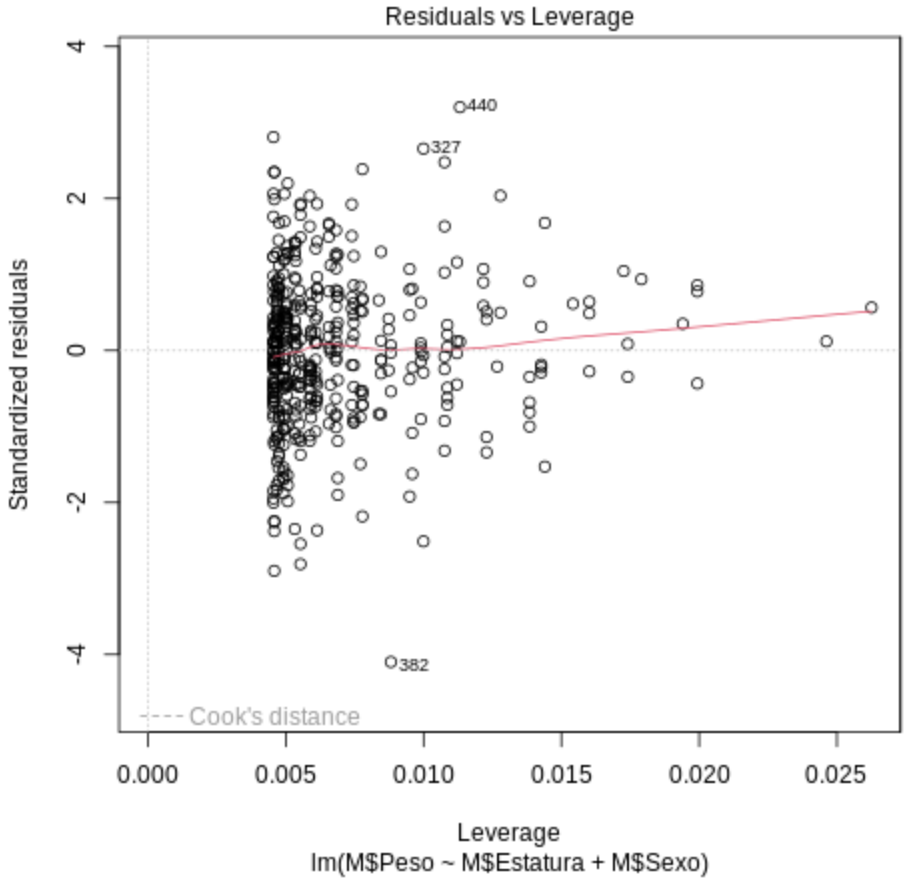
Generamos las gráficas de intervalos de confianza

In [ ]:
```
%%R
plot(A)
```


Residuals vs Fitted
Residuals
Fitted values
lm(M$Peso ~ M$Estatura + M$Sexo)
440
430
382
20
10
0
-10
-20
50
60
70
80


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(M$Peso ~ M$Estatura + M$Sexo)
440
430
382
4
2
0
-2
-4
-3
-2
-1
0
1
2
3


Scale-Location
382
440
430
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
50
60
70
80
Fitted values
lm(M$Peso ~ M$Estatura + M$Sexo)



2. Utiliza el comando: plot(modelo). Observa las gráficas obtenidas y contesta:
3. ***¿Cuáles son las diferencias y similitudes de estos gráficos con respecto a los que ya habías analizado?***

   La primer gráfica que muestra los residuos de los datos contra los datos útiles se asemeja bastante a la gráfica de homocedasticidad. En ambos casos podemos examinar como la distribución de los residuos es relativamente constante a través del rango de los valores ajustados. En ambas gráficas podemos examinar en cuál la dispersi´no de los residuos es relativamente constante a traves del rango de los valores ajustados. Si la dispersión varia sistemáticamente, entinces puede darse el caso de que existan problemas de heterocedasticidad. Además, ambas gráficas sirven para detectar patrones o tendencias en los residuos, o bien valores atípicos y puntos de inflexión.

   En cuanto a la QQPlot que arroja el comando plot(), podemos notar claramente que se trata de la misma gráfica que elaboramos para corroborar la normalidad de los datos en los puntos anteriores. De igual forma, la gráfica de dispersión es bastante similiar a la realizada para analizar los residuos, en específico para su homocedasticidad. Se comparan los datos ajustados contra los datos residuo, sin embargo, la gráfica utilizada en el punto anterior difiere en que se utiliza una línea recta que pasa por todos los puntos, en cambio esta gráfica posee una línea recta que va cambiando su dirección por cada valor ajustado que pasa.

2. ***Estos gráficos, ¿cambian en algo las conclusiones que ya habías obtenido?***
Las nuevas gráficas generadas mediante el comando plot(A) brindan una perspectiva visual más detallada de la relación entre los residuos y los valores ajustados, así como de la normalidad de los residuos. Sin embargo, estas gráficas no deberían cambiar sustancialmente las conclusiones previamente obtenidas, ya que están explorando los mismos aspectos de la calidad del modelo. Las tendencias, patrones y valores atípicos que se observen en estas gráficas podrían respaldar o reforzar las conclusiones anteriores sobre la validez del modelo, pero en sí no deberían cambiar la interpretación general. Es importante recordar que las gráficas son herramientas complementarias y deben ser consideradas junto con otros análisis y pruebas estadísticas para tomar decisiones informadas sobre la calidad del modelo.