Actividad11_Regresion_Lineal_Con_Interacción

Rodolfo Jesús Cruz Rebollar

2024-09-04

La recta de mejor ajuste

```
# Leer los datos del archivo csv

estatura_peso = read.csv("Estatura_Peso.csv")

head(estatura_peso)

##   Estatura  Peso Sexo
## 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
```

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

```
# Obtener la matriz de correlación de los datos

datos_M = subset(estatura_peso, estatura_peso$Sexo =="M")

datos_H = subset(estatura_peso, estatura_peso$Sexo =="H")

datos_general = data.frame(datos_H$Estatura,
                           datos_H$Peso, datos_M$Estatura, datos_M$Peso)

# Obtener matriz de correlación de los datos separados

cor(datos_general)

##                   datos_H.Estatura datos_H.Peso datos_M.Estatura
datos_M.Peso
## datos_H.Estatura     1.0000000000  0.846834792     0.0005540612
0.04724872
## datos_H.Peso         0.8468347920  1.000000000     0.0035132246
0.02154907
## datos_M.Estatura     0.0005540612  0.003513225     1.0000000000
0.52449621
## datos_M.Peso         0.0472487231  0.021549075     0.5244962115
1.00000000
```

Interpretación de la matriz de correlación

En la matriz de correlación obtenida previamente, es posible observar que existe un grado de correlación mayormente elevado entre la variable de la estatura de los hombres y el peso también de los hombres, siendo el nivel de correlación igual a 0.8468, por lo cual al ser un nivel de correlación positivo, eso indica que la estatura de los hombres es directamente proporcional al peso de los mismos, motivo por el cual, se espera que conforme la estatura de los hombres incrementa, también aumenta su peso. Además de lo anterior, en la matriz de correlación también se puede apreciar que en cuanto a las mujeres, la estatura de las mismas tiene la mayor correlación de todas con el peso de las mismas, siendo el grado de correlación entre ambas variables (estatura de mujeres y peso de mujeres) igual a 0.5244 que además al ser positivo, indica que se espera que entre mayor sea la estatura de las mujeres, su peso también será mayor, aunque dicha relación no es tan fuerte como en el caso de los hombres, por lo cual a partir de la matriz de correlación se concluye que en el caso de los hombres, la relación entre peso y estatura es mucho más fuerte que en el caso de las mujeres, por lo cual se afirma que el sexo de la persona, ya sea masculino, o femenino, tiene una influencia significativa en el peso que se espera que tenga dicha persona en función de su estatura, lo cual demuestra que los hombres son aquellas personas que ganan mucho mayor peso a medida que aumentan de estatura que las mujeres.

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

```
# Definir cantidad de variables para separar los datos principales

# 4 variables: estatura de hombres, peso de hombres, estatura de mujeres,
# peso de mujeres

n = 4

# Definir una matriz de valores nulos NA de n filas por 7 columnas para
# rellenarla con el resumen de las medidas estadísticas para cada una de
las
# las variables del conjunto de datos original

d = matrix(NA, ncol = 7, nrow = n)

# Ciclo for para calcular el resumen de medidas estadísticas por cada
variable del
# dataframe de datos original y por cada resumen calculado, guarde los
valores de los
# estadísticos en cada fila de la matriz d definida anteriormente

for(i in 1:n){

  # Por cada iteración, calcular el resumen de medidas estadísticas de
cada una de las
```

```
  # variables y guardar los valores de los estadísticos en cada fila i de
la matriz d
  # anterior

  d[i, ] = c(as.numeric(summary(datos_general[, i])), sd(datos_general[
,i]))

}

# Convertir la matriz d ya con las medidas estadísticas para las
diferentes variables
# en un dataframe para una mejor comprensión de los valores de los
estadísticos

medidas_estadisticas = as.data.frame(d)

# Definir nombres de las filas del dataframe creado

row.names(medidas_estadisticas)=c("H-Estatura","H-Peso","M-Estatura","M-
Peso")

# Definir los nombres de las columnas para el dataframe creado con
anterioridad

names(medidas_estadisticas)=c("Minimo","Q1","Mediana","Media","Q3","Máxim
o","Desv Est")

# Mostrar dataframe con valores de las medidas estadísticas por variable

medidas_estadisticas

##            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
```

La recta de mejor ajuste

Para encontrar la ecuación de regresión que representa el mejor ajuste para los datos en cuestión, a continuación se realizarán 3 modelos de regresión: uno que explique la relación entre peso y estatura de los hombres, otro que explique la relación entre peso y estatura en mujeres y un tercero explicando la relación entre peso y estatura de forma general entre ambos sexos de personas, tanto mujeres como hombres.

Modelo 1: relación peso y estatura de mujeres

```
# Modelo 1 que representa la relación entre el peso y la estatura en las
mujeres
```

```r
modelo.mujeres = lm(Peso ~ Estatura, datos_M)

# Mostrar el modelo que mejor representa el peso en función de la
estatura en las mujeres

modelo.mujeres

##
## Call:
## lm(formula = Peso ~ Estatura, data = datos_M)
##
## Coefficients:
## (Intercept)     Estatura
##      -72.56        81.15
```

Modelo 2: la relación entre peso y estatura en hombres

```r
# Modelo 2 para representar la relación entre el peso y la estatura en
los hombres

modelo.hombres = lm(Peso ~ Estatura, data = datos_H)

# Mostrar el modelo que mejor representa el peso en función de la
estatura en hombres

modelo.hombres

##
## Call:
## lm(formula = Peso ~ Estatura, data = datos_H)
##
## Coefficients:
## (Intercept)     Estatura
##      -83.68        94.66
```

Modelo 3: relación del peso con el sexo y la estatura tanto en hombres como en mujeres

```r
# Modelo 3 para explicar la relación del peso con el sexo y la estatura
de la persona
# ya sea hombre o mujer

modelo.HM = lm(Peso ~ Estatura + Sexo, estatura_peso)

modelo.HM

##
## Call:
## lm(formula = Peso ~ Estatura + Sexo, data = estatura_peso)
##
## Coefficients:
## (Intercept)     Estatura        SexoM
##      -74.75        89.26       -10.56
```

Ecuaciones de las rectas de mejor ajuste encontradas

Modelo 1: relación peso-estatura en hombres: $Peso = 94.66 * Estatura - 83.68$

Modelo 2: relación peso-estatura en mujeres: $Peso = 81.15 * Estatura - 72.56$

Modelo 3: relación del peso con la estatura y el sexo: $Peso = 89.26 * Estatura - 10.56 * Sexo - 74.75$

Summary del modelo peso vs estatura para las mujeres

```
# Resumen del modelo para las mujeres

summary(modelo.mujeres)

##
## Call:
## lm(formula = Peso ~ Estatura, data = datos_M)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3256  -4.1942   0.4004   4.2724  17.9114
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -72.560     14.041  -5.168 5.34e-07 ***
## Estatura      81.149      8.922   9.096  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.65 on 218 degrees of freedom
## Multiple R-squared:  0.2751, Adjusted R-squared:  0.2718
## F-statistic: 82.73 on 1 and 218 DF,  p-value: < 2.2e-16
```

Summary del modelo peso vs estatura para los hombres

```
# Resumen del modelo de regresión peso vs estatura para los hombres

summary(modelo.hombres)

##
## Call:
## lm(formula = Peso ~ Estatura, data = datos_H)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3881 -2.6073 -0.0665  2.4421 11.1883
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -83.685      6.663  -12.56   <2e-16 ***
## Estatura      94.660      4.027   23.51   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.678 on 218 degrees of freedom
## Multiple R-squared:  0.7171, Adjusted R-squared:  0.7158
## F-statistic: 552.7 on 1 and 218 DF,  p-value: < 2.2e-16
```

Summary del modelo peso vs estatura y sexo tanto en hombres como en mujeres

```
# Resumen del modelo de regresión del peso en función de la estatura y el
sexo

summary(modelo.HM)
```

```
##
## Call:
## lm(formula = Peso ~ Estatura + Sexo, data = estatura_peso)
##
## 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 ***
## Estatura     89.2604     4.5635  19.560   <2e-16 ***
## 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
```

Validación de los diferentes modelos generados

Prueba de Hipótesis para validar la significancia de los modelos de regresión:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Validación del modelo de regresión 1 (peso vs estatura en mujeres)

En cuanto al primer modelo de regresión lineal realizado para explicar la relación entre el peso y la estatura en las mujeres, es posible observar en el summary de dicho modelo que el coeficiente β_1 del mismo posee un valor de significancia inferior a 2e-16, lo cual al ser mucho menor que el nivel de significancia de 0.03 para la prueba, se cuenta con suficiente evidencia estadística para rechazar H_0, motivo por el cual, se afirma que el coeficiente β_1 sí es significativo en el contexto estadístico. Adicionalmente, también cabe mencionar que en cuanto a la significancia del modelo en general, éste modelo posee un valor p inferior a 2.2e-16, lo cual al ser menor que 0.03 se tiene evidencia estadística suficiente para rechazar H_0, lo cual nos lleva a afirmar que el modelo sí es estadísticamente significativo. Además, en relación al porcentaje de variación explicada por el modelo (R^2), es posible observar en el summary del modelo que el coeficiente R^2 del modelo es igual a 0.2718, lo cual señala que el modelo es capaz de explicar el 27.18% de la variabilidad total presente en los datos de mujeres para estatura y peso, por lo cual, se puede concluir que aunque estadísticamente hablando el modelo sea significativo, éste mismo no es capaz de explicar un porcentaje de variabilidad suficiente de los datos en cuestión, por lo que las predicciones derivadas de dicho modelo no serán suficientemente confiables como para reflejar la relación verdadera entre el peso y la estatura en la población femenina.

Validación del modelo de regresión 2 (peso vs estatura en hombres)

En cuanto al segundo modelo de regresión lineal realizado para explicar la relación entre el peso y la estatura en los hombres, es posible observar en el summary de dicho modelo que el coeficiente β_1 del mismo posee un valor de significancia inferior a 2e-16, lo cual al ser mucho menor que el nivel de significancia de 0.03 para la prueba, se cuenta con suficiente evidencia estadística para rechazar H_0, motivo por el cual, se afirma que el coeficiente β_1 sí es significativo en el contexto estadístico. Adicionalmente, también cabe mencionar que en cuanto a la significancia del modelo en general, éste modelo posee un valor p inferior a 2.2e-16, lo cual al ser menor que 0.03 se tiene evidencia estadística suficiente para rechazar H_0, lo cual nos lleva a afirmar que el modelo sí es estadísticamente significativo. Además, en relación al porcentaje de variación explicada por el modelo (R^2), es posible observar en el summary del modelo que el coeficiente R^2 del modelo es igual a 0.7158, lo cual señala que el modelo es capaz de explicar el 71.58% de la variabilidad total presente en los datos de hombres para estatura y peso, por lo cual, se puede concluir que además de que estadísticamente hablando el modelo sea significativo, éste mismo también es capaz de explicar un porcentaje de variabilidad suficientemente aceptable de los datos en cuestión, por lo que las predicciones derivadas de dicho modelo serán en su mayoría suficientemente confiables como para reflejar la relación verdadera de la estatura con el peso en la población masculina.

Validación del modelo de regresión 3 (peso vs estatura y sexo)

En cuanto al primer modelo de regresión lineal realizado para explicar la relación del peso con la estatura y el sexo de las personas, es posible observar en el summary de dicho modelo que los coeficientes β_i del mismo poseen un valor de significancia inferior a 2e-16, lo cual al ser mucho menor que el nivel de significancia de 0.03 para la prueba, se cuenta con suficiente evidencia estadística para rechazar H_0, motivo por el cual, se afirma que los coeficientes β_i sí son significativos en el contexto estadístico. Adicionalmente, también cabe mencionar que en cuanto a la significancia del modelo en general, éste modelo posee un valor p inferior a 2.2e-16, lo cual al ser menor que 0.03 se tiene evidencia estadística suficiente para rechazar H_0, lo cual nos lleva a afirmar que el modelo sí es estadísticamente significativo. Además, en relación al porcentaje de variación explicada por el modelo (R^2), es posible observar en el summary del modelo que el coeficiente R^2 del modelo es igual a 0.7827, lo cual señala que el modelo es capaz de explicar el 78.27% de la variabilidad total presente en los datos de ambos sexos de personas para estatura y peso, por lo cual, se puede concluir que además de que estadísticamente hablando el modelo sea significativo, éste mismo es capaz de explicar un porcentaje de variabilidad suficiente de los datos en cuestión, incluso mayor que el porcentaje de variabilidad explicado por el segundo modelo (peso vs estatura en hombres), por lo que las predicciones derivadas de dicho modelo tendrá incluso una mayor confiabilidad que las predicciones resultantes del segundo modelo, por lo cual lograrán reflejar de una forma mayormente confiable, la relación verdadera entre del sexo y la estatura con el peso en la población general.

Gráfico de dispersión de los datos y de la recta de mejor ajuste

Gráfico de dispersión y recta de mejor ajuste para las mujeres

```
# Gráfico de dispersión de peso contra estatura de los datos 
correspondientes a
# las mujeres

plot(datos_M$Estatura, datos_M$Peso, col = "red",
     main = "Mujeres: Peso vs Estatura", xlab = "Estatura", ylab = 
"Peso")

abline(modelo.mujeres, col = "darkgreen", lwd = 3)
```

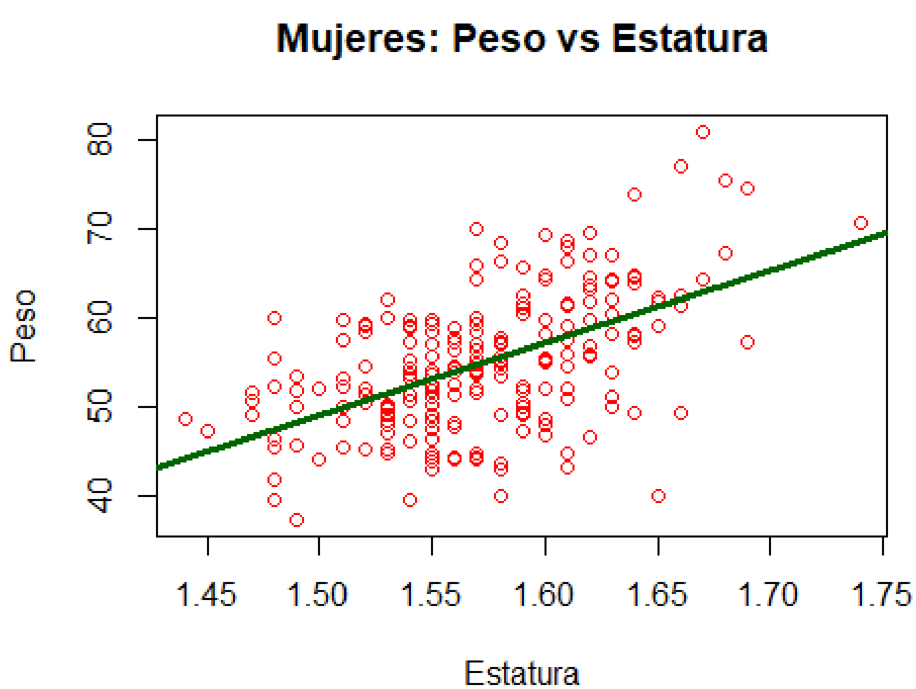


Gráfico de dispersión y recta de mejor ajuste para los hombres

```
# Gráfico de dispersión de peso contra estatura de los datos 
correspondientes a
```

```r
# Los hombres

plot(datos_H$Estatura, datos_H$Peso, col = "blue",
     main = "Hombres: Peso vs Estatura", xlab = "Estatura", ylab =
"Peso",
     pch = 19)

abline(modelo.hombres, col = "darkred", lwd = 2.5)
```

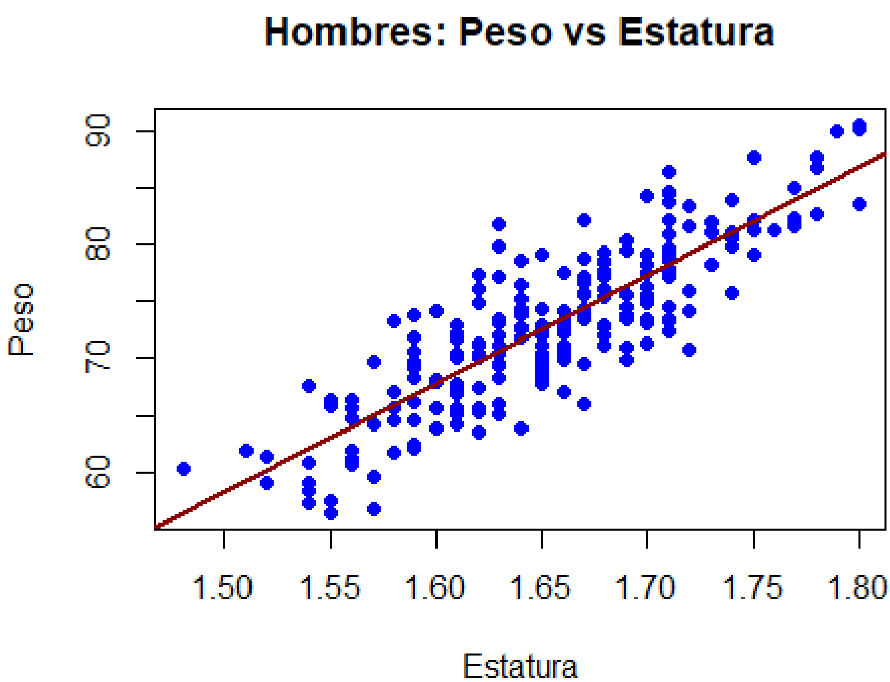
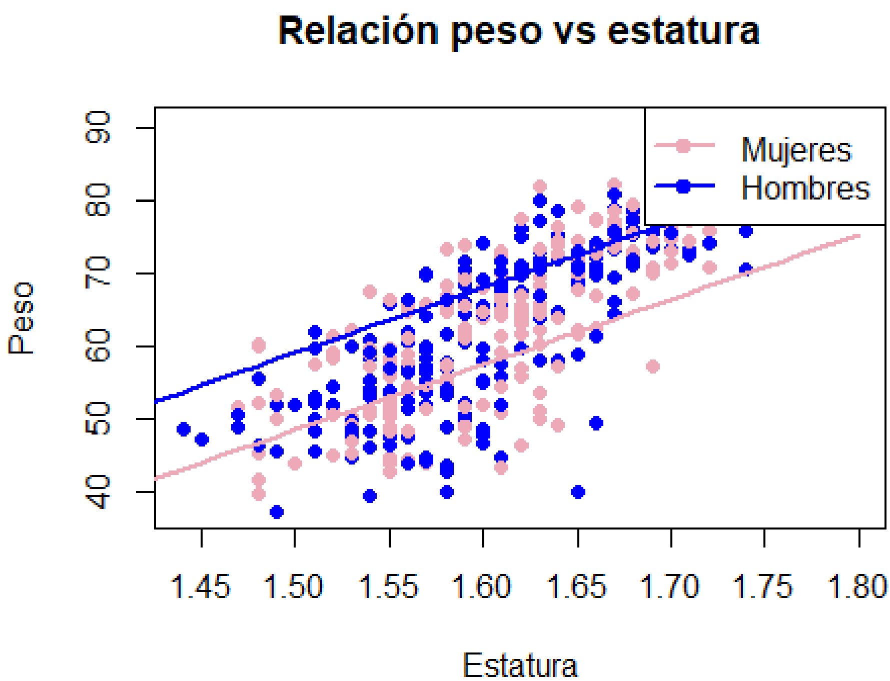


Gráfico de dispersión y recta de mejor ajuste para peso en función de sexo y estatura

```r
# Separar los coeficientes del modelo en 3 variables por separado

b0 = modelo.HM$coefficients[1] # intercepto del modelo beta0
b1 = modelo.HM$coefficients[2] # coeficiente para la estatura beta1
b2 = modelo.HM$coefficients[3] # coeficiente para el sexo beta2

# Definir función para predecir el peso en función del sexo y estatura
para las mujeres

Ym = function(x){b0 + b2 + b1 * x}

# Definir función para predecir el peso en función del sexo y estatura
para las mujeres

Yh = function(x){b0 + b1 * x}
```

```
# Colores para graficar los puntos y las rectas correspondientes a
mujeres y hombres
# rosado: mujeres, azul: hombres

colores = c("pink2", "blue")

# Graficar la estatura de la población general vs su peso con los colores
indicados para
# distinguir los datos correspondientes a las mujeres y a los hombres

plot(estatura_peso$Estatura, estatura_peso$Peso, col = colores, pch = 19,
ylab="Peso",
     xlab="Estatura", main="Relación peso vs estatura")

# Datos de estatura desde la estatura mínima hasta la máxima con
incrementos de 0.01

x = seq(1.40, 1.80, 0.01)

# Graficar recta estatura contra peso predicho por el modelo para las
mujeres

lines(x, Ym(x), col = "pink2", lwd = 2)

# Graficar recta de estatura contra peso predicho por el modelo para
hombres

lines(x, Yh(x), col = "blue", lwd = 2)

# Agregar una leyenda al gráfico para distinguir fácilmente entre hombres
y mujeres

legend("topright", legend = c("Mujeres", "Hombres"), col =
c("pink2","blue"), lwd = 2,
       pch = 19)
```



Interpreta en el contexto del problema cada uno de los análisis que hiciste.

En primera instancia, en el primer análisis realizado referente al modelo de peso en función de la estatura exclusivamente en los hombres, cabe mencionar que en el contexto del problema, a medida que se incrementa la estatura en los hombres, también se incrementa de forma gradual su peso, lo cual se evidencia en el gráfico de dispersión de los datos y de la recta del modelo para los datos de los hombres, mientras que de forma similar, aunque con una menor fuerza de correlación, las mujeres también tienen una tendencia a experimentar un aumento en su peso a medida que su estatura también incrementa, aunque en el caso particular de las mujeres, se observa que hay algunos datos que se encuentran más alejados de la línea principal conforme la estatura aumenta, lo cual respalda el hecho de que el grado de correlación entre estatura y peso en el caso de las mujeres sea menor que en el caso de los hombres. Por otro lado, en cuanto al modelo que explica la relación entre el peso en función de la estatura y el sexo de la persona, es posible interpretar que la probabilidad de que una persona tenga un mayor peso se incrementa significativamente cuando la persona es de sexo masculino (hombres), mientras que si es mujer, dicha probabilidad disminuye, lo cual tiene sentido, ya que los hombres tienden a realizar actividades que implican un mayor grado de esfuerzo, mientras que las mujeres no, motivo por el cual, los hombres también tienen la tendencia a tener una dieta más o menos balanceada en función del grado de intensidad de la actividad física que realicen, esto con el proósito de contar con los nutrientes necesarios y suficientes para poder llevarlas a cabo de una forma saludable y para que a su vez, su rendimiento en ellas sea el mejor posible.

Interpreta en el contexto del problema

¿Qué información proporciona β_0 sobre la relación entre la estatura y el peso de hombres y mujeres?

En cuanto a los modelos de regresión lineal realizados, el parámetro del modelo β_0 que corresponde al valor del intercepto del mismo, hace refererencia al valor que adopta la variable dependiente del modelo cuando el valor de la variable independiente del modelo es igual a 0, lo cual aplicado a este problema en específico, indica que hipotéticamente hablando, cuando la estatura de una persona es igual a 0, lo cual no posee ningún sentido físico, el peso de dicha persona será igual a -74.75 lo cual tampoco tiene sentido en un contexto físico, ya que tanto el peso como la estatura no pueden ser negativos, o iguales a 0, motivo por el cual, será necesario seguir generando posibles modelos para explicar de una mejor forma los datos en cuestión, principalmente quitando aquellos factores o variables que más allá de resultar ser estadísticamente significativas, no tengan ningún sentido en el contexto del problema, y después de quitar esas variables una por una, volver a generar otro modelo para explicar los datos y ver de qué manera mejoran los aspectos generales del modelo, tales como la significancia de dicho modelo en general, la significancia de los coeficientes correspondientes a cada una de las variables predictoras, entre otros aspectos relevantes del modelo, asegurando que al mismo tiempo, cada uno de los coeficientes del modelo, incluyendo su intercepto β_0, sí tenga sentido en el contexto del problema que se esté intentando resolver.

¿Cómo interpretas β_1 en la relación entre la estatura y el peso de hombres y mujeres?

Adicionalmente, en cuanto al coeficiente β_1 en el contexto de la relación entre la estatura de las personas y el peso de las mismas, dicho coeficiente del modelo hace referencia al hecho de que por cada unidad que cambie el valor de la estatura de una persona, el peso de la misma cambiará β_1 cantidad de unidades, es decir, que por cada metro que incremente la estatura de una persona, el incremento en su peso será de la cantidad de unidades especificada por dicho coeficiente β_1, por ejemplo, para el caso particular de las mujeres, por cada metro que aumenten de estatura, su peso experimentará un incremento de 81.15 kilogramos, mientras que por otro lado, en el caso específico de los hombres, por cada metro que aumenten de estatura, su peso experimentará un aumento de 94.66 kilogramos.

Nuevo modelo: interacción de estatura con el sexo

Obtención del modelo e interpretación de variables dummy

```
# Generar nuevo modelo esta vez tomando en cuenta la interacción de
estatura y sexo

new.model = lm(Peso ~ Estatura * Sexo, data = estatura_peso)
```

```r
# Mostrar el nuevo modelo relacionando el peso con la interacción entre estatura y sexo

new.model

##
## Call:
## lm(formula = Peso ~ Estatura * Sexo, data = estatura_peso)
##
## Coefficients:
##    (Intercept)        Estatura           SexoM  Estatura:SexoM
##         -83.68           94.66           11.12          -13.51

# Mostrar el resumen o summary del nuevo modelo

summary(new.model)

##
## Call:
## lm(formula = Peso ~ Estatura * Sexo, data = estatura_peso)
##
## 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 ***
## Estatura          94.660      5.882  16.092   <2e-16 ***
## SexoM             11.124     14.950   0.744    0.457
## Estatura: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
```

En cuanto a las variables dummy, es posible afirmar que éstas mismas hace referencia al sexo de las personas, en el sentido de que la variable sexo se puede subdividir en dos variable adicionales llamadas variables dummy, la primera de ellas es la variable dummy referente a los hombres, mientras que la segunda es la variable dummy que hace alusión a las mujeres, motivo por el cual, es posible interpretar que en cuanto a la variable dummy de hombres, esta misma indica simplemente si la persona en cuestión es hombre o no, con un 1 en caso de que sí lo sea, o un 0 en caso contrario, mientras que de forma similar para el caso de la variable dummy de las mujeres, el valor 1 significa que una persona en particular es mujer, mientras que el 0 significa que no lo es, en otras palabras, en la variable dummy de hombres 1 significa que la persona es hombre y 0 significa que es mujer y de forma similar, para el caso de la variable dummy de mujeres, 1 significa que la persona es mujer y 0 que es hombre.

Significancia del modelo

Significancia del modelo con α de 0.03

Prueba de hipótesis:

H_0: El modelo generado no es estadísticamente significativo.

H_1: el modelo generado sí es estadísticamente significativo.

```
# Mostrar resumen del nuevo modelo con valores de significancia

summary(new.model)

##
## Call:
## lm(formula = Peso ~ Estatura * Sexo, data = estatura_peso)
##
## 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 ***
## Estatura         94.660      5.882  16.092   <2e-16 ***
## SexoM            11.124     14.950   0.744    0.457
## Estatura: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
```

En el summary del nuevo modelo se aprecia que el p valor del mismo es inferior a 2.2e-16, lo cual resulta ser mucho menor que 0.03, motivo por el cual, se cuenta con evidencia estadística suficiente para rechazar H_0, por lo tanto es posible concluir que el modelo generado sí es estadísticamente significativo.

Significancia de β_i

Prueba de hipótesis:

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

```
# Nuevamente, mostrar el resumen del nuevo modelo para verificar la
significancia de
# sus coeficientes
```

```r
summary(new.model)

##
## Call:
## lm(formula = Peso ~ Estatura * Sexo, data = estatura_peso)
##
## 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 ***
## Estatura         94.660      5.882  16.092   <2e-16 ***
## SexoM            11.124     14.950   0.744    0.457
## Estatura: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
```

De acuerdo con lo mostrado en el resumen del nuevo modelo generado, se observa que el intercepto del modelo β_0 posee un p valor inferior a 2e-16, lo cual al ser menor que 0.03, se tiene evidencia estadística suficiente para rechazar H_0, por lo que en este caso, se concluye que el coeficiente β_0 sí es estadísticamente significativo. Por otro lado, en cuanto al coeficiente β_1 asociado a la estatura de las personas, se aprecia que su valor p es inferior a 2e-16, por lo cual en este caso particular se tiene evidencia estadística para rechazar H_0, motivo por el cual, se concluye que el coeficiente β_1 sí es estadísticamente significativo. Además de lo anterior, en cuanto al coeficiente β_2 asociado al sexo de las personas, se aprecia que dicho coeficiente posee un valor p de 0.457, lo cual es mayor a 0.03, motivo por el cual en ese caso se tiene evidencia estadística para no rechazar H_0, por lo que se concluye que β_2 no es estadísticamente significativo para el modelo. Adicionalmente para el coeficiente β_3 asociado con la interacción entre el sexo de las personas y su estatura, se observa que para dicho coeficiente el valor p es de 0.147, lo cual es mayor a 0.03, por lo que en este caso se tiene evidencia estadística para no rechazar H_0, por lo tanto, se concuye que el coeficiente β_3 no es estadísticamente significativo para el modelo.

Porcentaje de variación explicada por el modelo

```r
# Mostrar resumen del nuevo modelo para visualizar el valor de los coeficientes r^{2}

summary(new.model)

##
## Call:
```

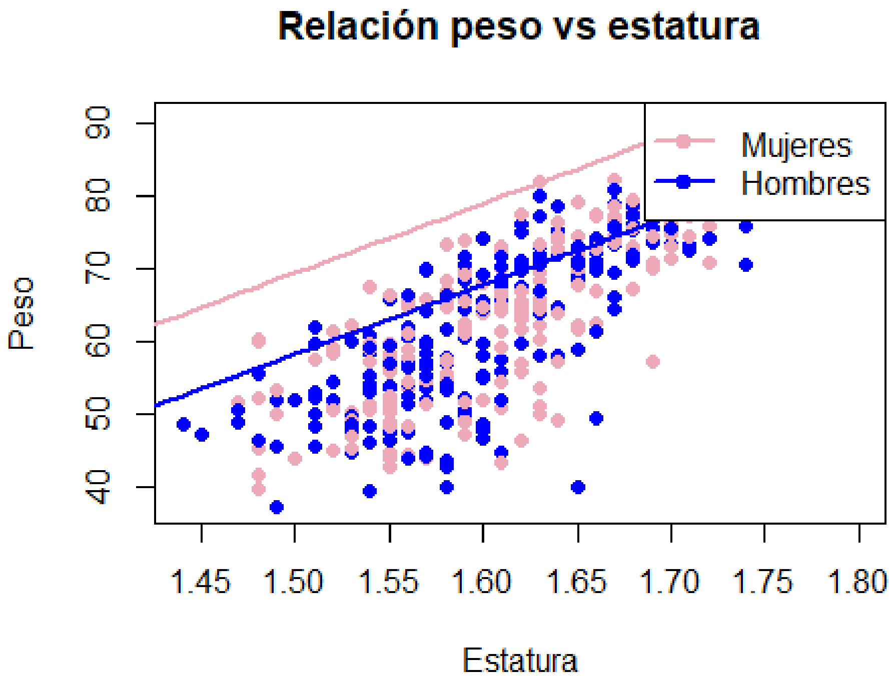
```
## lm(formula = Peso ~ Estatura * Sexo, data = estatura_peso)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3256  -3.1107   0.0204   3.2691  17.9114
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -83.685      9.735  -8.597   <2e-16 ***
## Estatura           94.660      5.882  16.092   <2e-16 ***
## SexoM              11.124     14.950   0.744    0.457
## Estatura: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
```

En el resumen del modelo se puede observar que el valor que adopta el coeficiente R^2 ajustado para varias variables predictoras es igual a 0.7832, lo cual significa que el modelo es capaz de explicar el 78.32% de la variabilidad total de los datos involucrados, por lo cual solamente resta un 21.68% de dicha variabilidad sin explicar por el modelo, por lo cual se puede afirmar que en términos generales, el modelo es capaz de explicar la mayoría de la variabilidad de los datos originales y por lo tanto, las predicciones derivadas del mismo también serán mayormente confiables.

Diagrama de dispersión de los datos y la recta de mejor ajuste

```r
# Separar los coeficientes del modelo en 3 variables por separado

b0_new = new.model$coefficients[1] # intercepto del modelo beta0
b1_new = new.model$coefficients[2] # coeficiente para la estatura beta1
b2_new = new.model$coefficients[3] # coeficiente para el sexo beta2

# Definir función para predecir el peso en función del sexo y estatura
para las mujeres

Ym = function(x){b0_new + b2_new + b1_new * x}

# Definir función para predecir el peso en función del sexo y estatura
para los hombres

Yh = function(x){b0_new + b1_new * x}

# Colores para graficar los puntos y las rectas correspondientes a
mujeres y hombres
# rosado: mujeres, azul: hombres

colores = c("pink2", "blue")
```

```
# Graficar la estatura de la población general vs su peso con los colores indicados para
# distinguir los datos correspondientes a las mujeres y a los hombres

plot(estatura_peso$Estatura, estatura_peso$Peso, col = colores, pch = 19, ylab="Peso",
     xlab="Estatura", main="Relación peso vs estatura")

# Datos de estatura desde la estatura mínima hasta la máxima con incrementos de 0.01

x = seq(1.40, 1.80, 0.01)

# Graficar recta estatura contra peso predicho por el modelo para las mujeres

lines(x, Ym(x), col = "pink2", lwd = 2)

# Graficar recta de estatura contra peso predicho por el modelo para hombres

lines(x, Yh(x), col = "blue", lwd = 2)

# Agregar una leyenda al gráfico para distinguir fácilmente entre hombres y mujeres

legend("topright", legend = c("Mujeres", "Hombres"), col = c("pink2","blue"), lwd = 2,
       pch = 19)
```



En conclusión, en cuanto al análisis de la significancia del modelo y sus coeficientes, es posible concluir que el modelo sí fue significativo, lo cual en el contexto del problema significa que el modelo sí es funcional para predecir el peso de las personas en base a su estatura y su sexo, mientras que de forma similar, al concluir que la variable de estatura es estadísticamente significativa, en el contexto del problema significa que la estatura sí tiene una influencia en el peso esperado de las personas, dado que entre mayor sea la estatura de las personas, su cuerpo será igualmente más grande, por lo cual es espera que su peso también aumente, además de lo anterior, la variable de sexo junto con la interacción entre estatura y sexo resultaron ser no significativas, lo cual en el contexto del problema significa que el sexo no posee una influencia significativa en el peso de las personas, además de que también las variables de estatura y de sexo no influyen de forma significativa entre ellas, por lo que dicha interacción no es estadísticamente significativa para el modelo de regresión, además se observa que los puntos del gráfico correspondientes a las mujeres no se ajustan de forma adecuada a la recta color rosa asociada a las mujeres, mientras que los puntos azules correspondientes a los hombres sí se ajustan en su mayoría de forma adecuada a la recta color azul igualmente asociada a los datos de los hombres, lo cual nos indica que el modelo que incluye la interacción ente sexo y estatura no es adecuado para explicar el peso de las personas en base a dicha interacción, lo cual respalda el hecho de que estadísticamente hablando, la interacción entre sexo y estatura haya resultado no ser significativa.

Interpreta en el contexto del problema:

¿Qué información proporciona β_0 sobre la relación entre la estatura y el peso de hombres y mujeres? Interpreta y compara entre este modelo con los 3 modelos anteriores.

En general, el coeficiente β_0 representa el valor base del modelo, en este caso el valor base del peso de las personas, mismo que ocurre cuando todas las variables predictoras, en este caso estatura y sexo son iguales a 0, por lo cual en el contexto del problema, esto hace referencia a que cuando la estatura de una persona es igual a 0 (lo cual no tiene sentido) y además, dicha persona no es mujer sino hombre (sexoM = 0), el peso predicho para esa persona es igual a -83.68, lo cual no tiene sentido, ya que el peso no puede tomar valores negativos, por lo cual, comparando los valores del coeficiente β_0 de los 4 modelos obtenidos, se tiene que dicho coeficiente tiene un valor de -83.68 en el modelo nuevo, -74.75 en el modelo que contempla personas de ambos sexos, de -83.68 en hombres y -72.56 en mujeres.

¿Cómo interpretas β_i en la relación entre la estatura y el peso de hombres y mujeres? Interpreta y compara entre este modelo con los 3 modelos anteriores

En términos generales, el coeficiente β_1 se refiere a la cantidad de unidades que cambia el peso a medida que cambia la estatura de las personas, por lo cual en el caso del nuevo modelo, dicho coeficiente tiene un valor de 94.66, mientras que en el modelo de ambos sexos dicho coeficiente tiene valor de 89.26, mientras que en el modelo para hombres su valor es de 94.66 y en el modelo de mujeres su valor es igual a 81.15, por lo cual, se aprecia que dicho coeficiente tiene su mayor valor en el modelo que considera la interacción entre sexo y estatura, además del otro modelo que solamente toma en cuenta la estatura de los hombres para predecir el peso de los mismos.

Indica cuál(es) de los modelos probados para la relación entre peso y estatura entre hombres y mujeres consideras que es más apropiado y explica por qué.

En conclusión, el modelo más apropiado que explica de mejor manera la relación entre peso y estatura tanto en hombres como en mujeres es aquel que considera la ocurrencia ya sea de la estatura o del peso, en el sentido de que dichos factores predictores no interactúen entre ellos, dado que en el caso del modelo que contempla ambos sexos de personas y no la interacción entre sexo y estatura, se puede apreciar que todos los coeficientes β_i de dicho modelo son significativos estadísticamente hablando, lo cual significa que todas las variables predictoras tienen una influencia significativa en el peso de las personas, por lo cual en el caso de dicho modelo, se cumple el hecho de que todas sus variables independientes sean estadísticamente significativas, mientras que en el modelo donde se toma en cuenta la interacción entre sexo y estatura no todas las variables predictoras son significativas, por lo que en el caso de este segundo modelo, las variables no significativas ocasionan que el rendimiento del modelo que contempla la interacción entre sexo y estatura no sea el más adecuado, por lo que el modelo ideal es el que considera tanto el sexo como la estatura de las personas pero sin que dichos factores interactúen entre ellos.

Validez de los modelos

1. Normalidad de residuos

H_0: los datos provienen de una población normal.

H_1: los datos no provienen de una población normal.

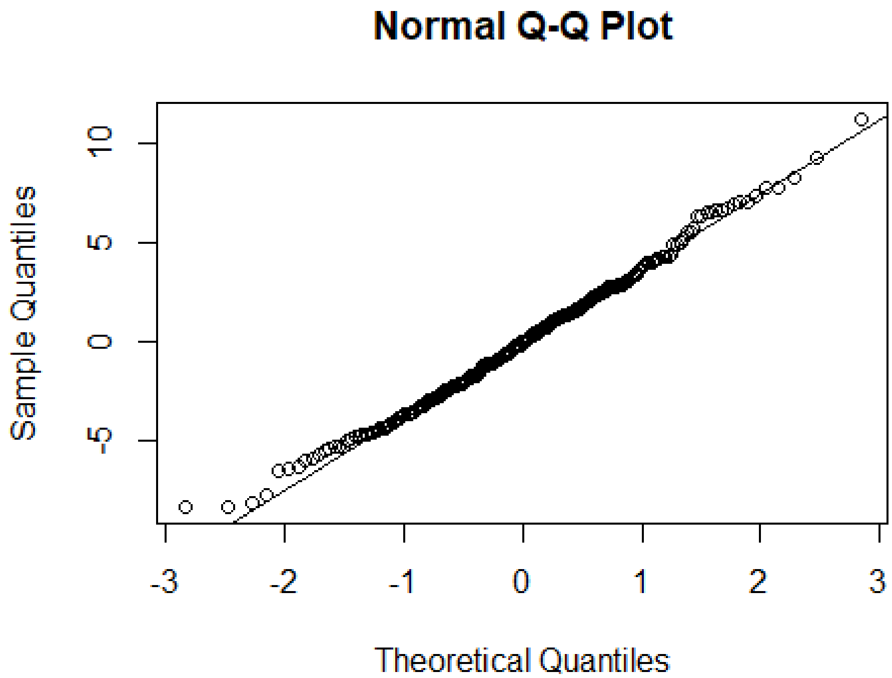
```
# libreria para tests de normalidad

library(nortest)
```

Modelo 1: hombres

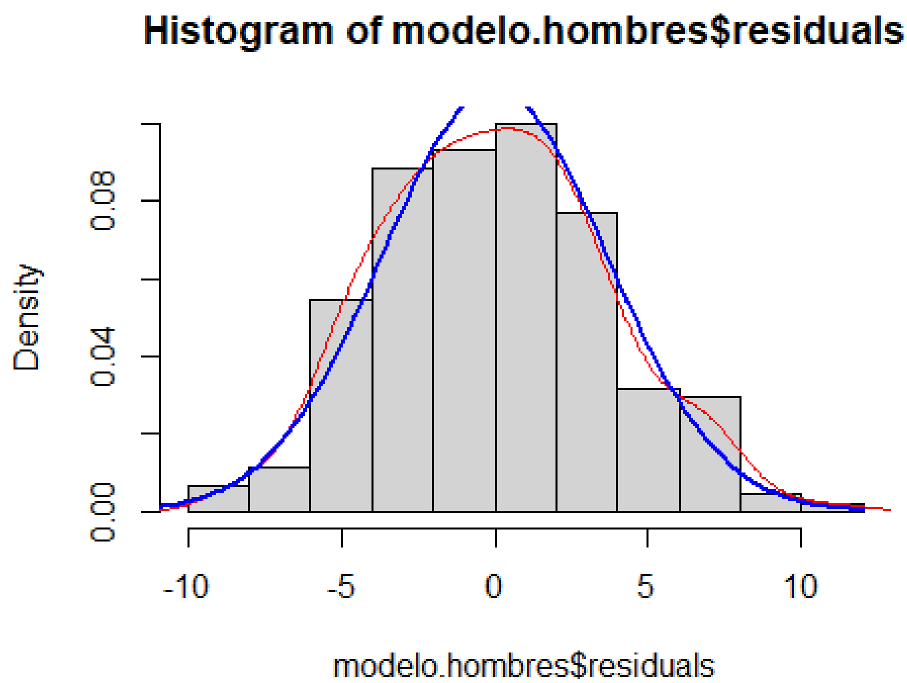
```
# Realizar test de normalidad de Anderson Darling para residuos del
modelo 1 para hombres

ad.test(modelo.hombres$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  modelo.hombres$residuals
## A = 0.3009, p-value = 0.5771
```

```
# Graficar el QQplot para ver el porqué se rechaza o se acepta normalidad
de residuos en
# el test de Anderson Darling

qqnorm(modelo.hombres$residuals)
qqline(modelo.hombres$residuals)
```

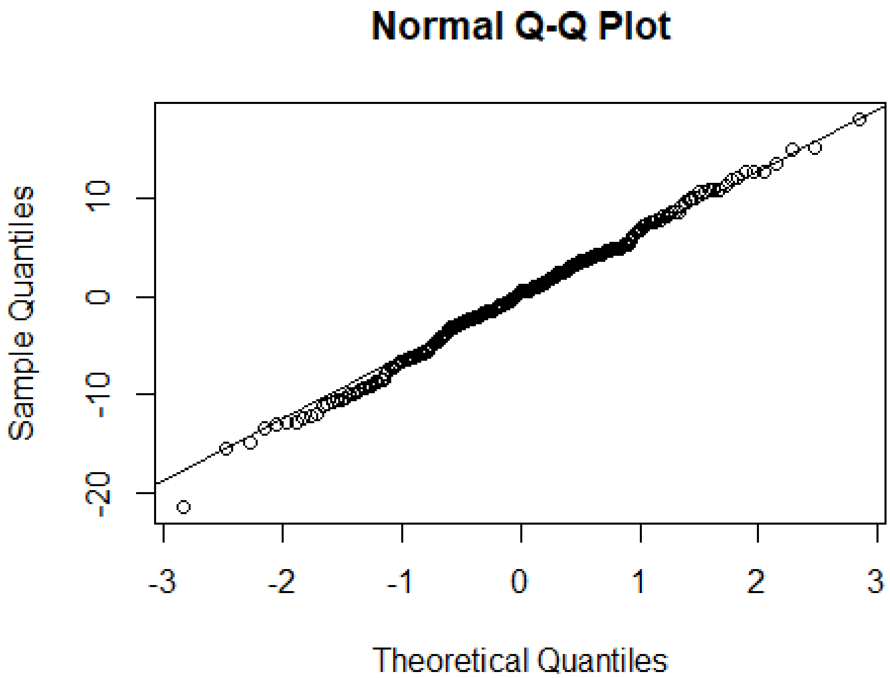



```
hist(modelo.hombres$residuals, freq=FALSE)
lines(density(modelo.hombres$residuals), col="red")
curve(dnorm(x, mean = mean(modelo.hombres$residuals), sd =
sd(modelo.hombres$residuals)),
      from=-12, to=12, add=TRUE, col="blue",lwd=2)
```

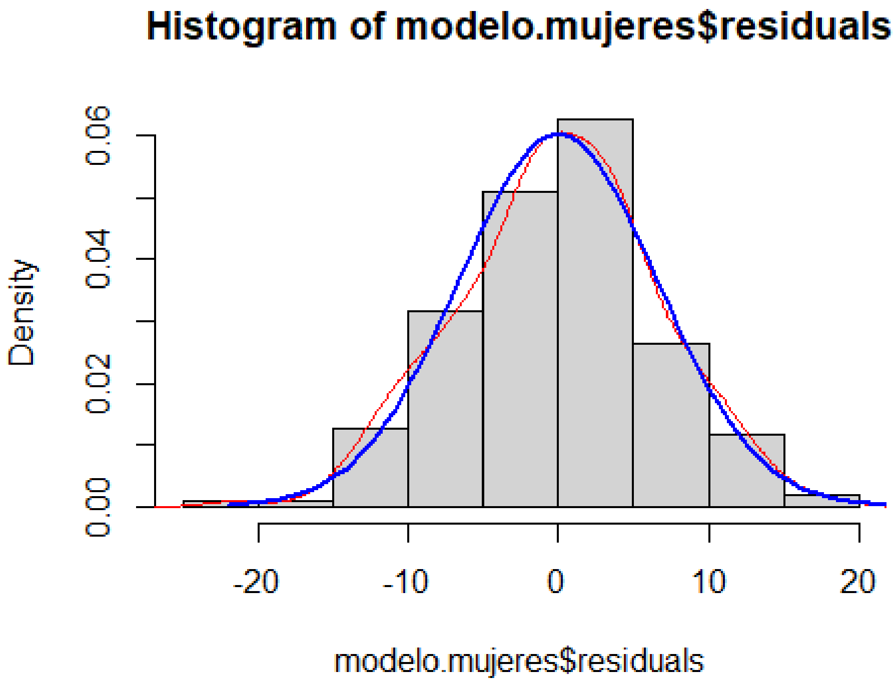


Modelo 2: mujeres

```
# Realizar test de normalidad de Anderson Darling para residuos del modelo 2 para mujeres

ad.test(modelo.mujeres$residuals)

##
##  Anderson-Darling normality test
##
## data:  modelo.mujeres$residuals
## A = 0.24899, p-value = 0.7451

# Graficar el QQplot para ver el porqué se rechaza o se acepta normalidad de residuos en
# el test de Anderson Darling

qqnorm(modelo.mujeres$residuals)
qqline(modelo.mujeres$residuals)
```

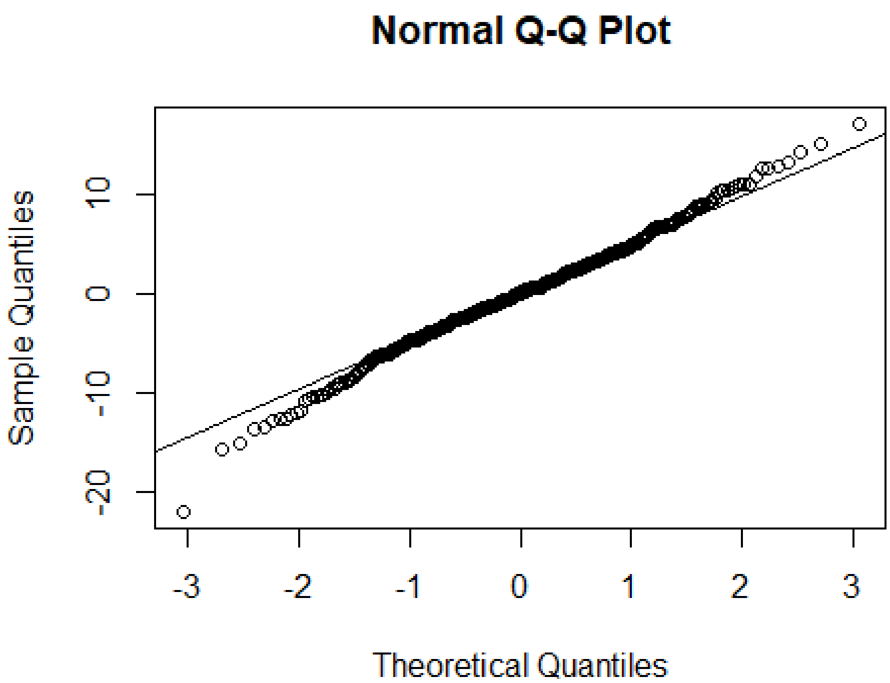



```
hist(modelo.mujeres$residuals, freq=FALSE)
lines(density(modelo.mujeres$residuals), col="red")
curve(dnorm(x, mean = mean(modelo.mujeres$residuals), sd =
sd(modelo.mujeres$residuals)),
      from=-22, to=22, add=TRUE, col="blue",lwd=2)
```

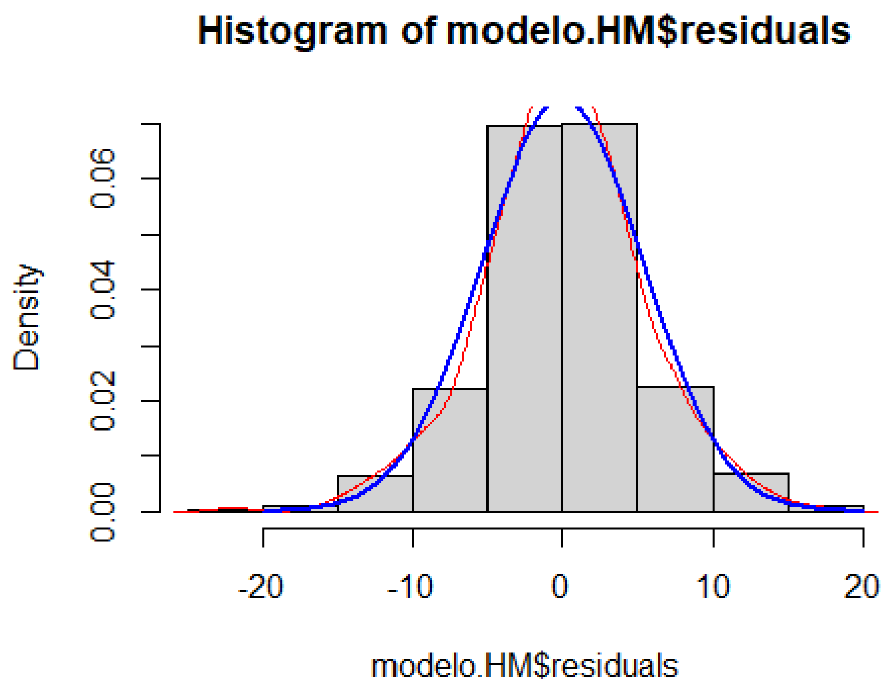


Modelo 3: hombres y mujeres

```
# Realizar test de normalidad de Anderson Darling para residuos del
modelo 3 para ambos
# sexos

ad.test(modelo.HM$residuals)

##
##  Anderson-Darling normality test
##
## data:  modelo.HM$residuals
## A = 0.79651, p-value = 0.03879

# Graficar el QQplot para ver el porqué se rechaza o se acepta normalidad
de residuos en
# el test de Anderson Darling

qqnorm(modelo.HM$residuals)
qqline(modelo.HM$residuals)
```

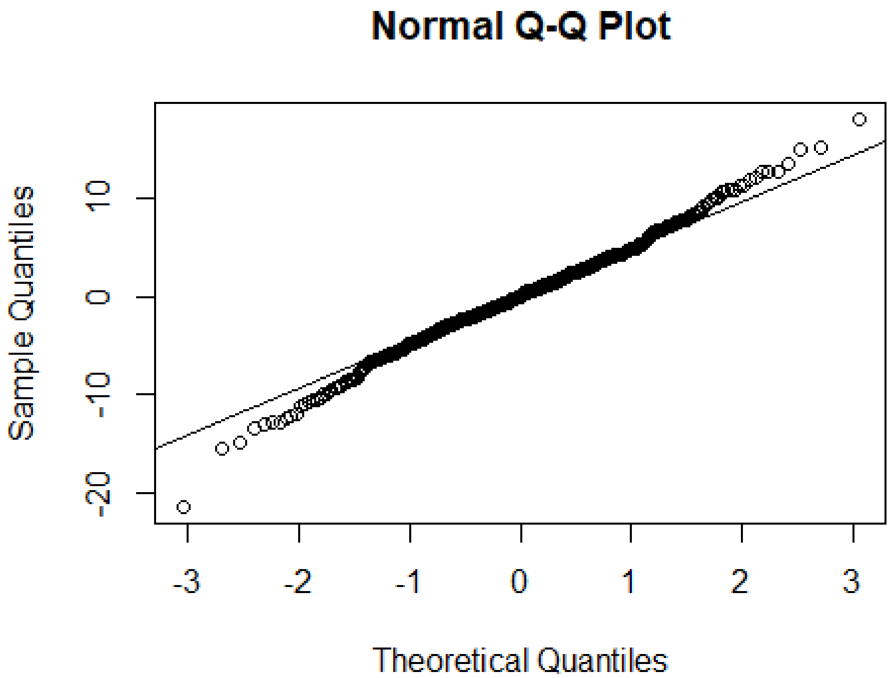



```
hist(modelo.HM$residuals, freq=FALSE)
lines(density(modelo.HM$residuals), col="red")
curve(dnorm(x, mean = mean(modelo.HM$residuals), sd =
sd(modelo.HM$residuals)),
      from=-20, to=20, add=TRUE, col="blue",lwd=2)
```

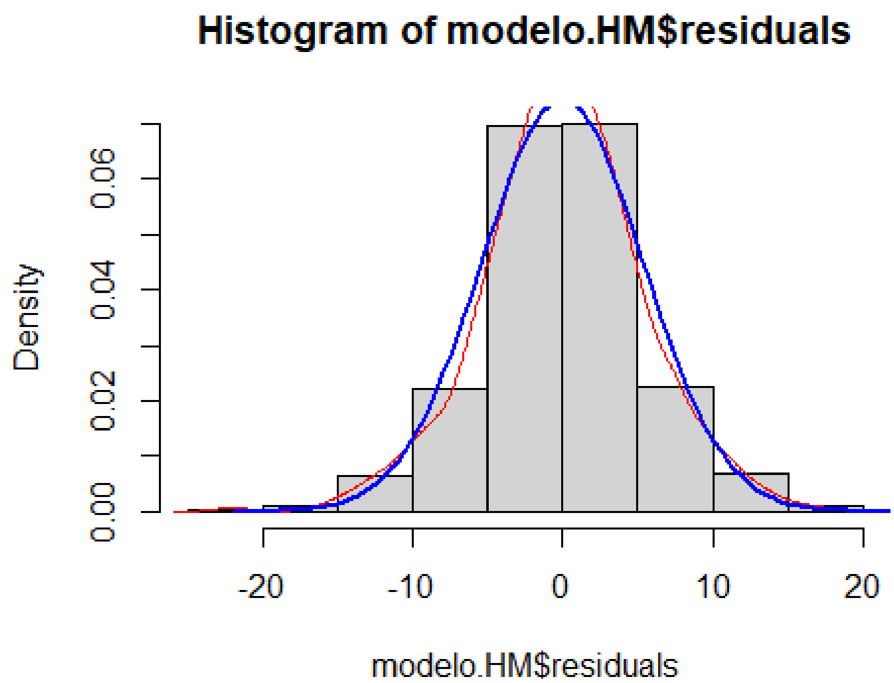


Modelo 4: Interacción entre sexo y estatura

```
# Realizar test de normalidad de Anderson Darling para residuos del
modelo 4 para la
# interacción entre el sexo y la estatura

ad.test(new.model$residuals)

##
##  Anderson-Darling normality test
##
## data:  new.model$residuals
## A = 0.8138, p-value = 0.03516

# Graficar el QQplot para ver el porqué se rechaza o se acepta normalidad
de residuos en
# el test de Anderson Darling

qqnorm(new.model$residuals)
qqline(new.model$residuals)
```




```
hist(modelo.HM$residuals, freq=FALSE)
lines(density(modelo.HM$residuals), col="red")
curve(dnorm(x, mean = mean(modelo.HM$residuals), sd =
sd(modelo.HM$residuals)),
      from=-22, to=22, add=TRUE, col="blue",lwd=2)
```



En resumen, se aprecia que para el modelo 1 (solo hombres), el valor p de la prueba de hipótesis es de 0.5771, lo cual es mayor que 0.05, por lo que no se rechaza H_0, lo cual significa que se puede concluir que los datos de dicho modelo provienen de una población normal, además, en el qqplot se observa que a pesar de haber ligeros alejamientos de los puntos del gráfico con respecto a la recta de normalidad, la gran mayoría de ellos se unican sobre dicha recta, respaldando el resultado del test de Anderson Darling. Además, para el modelo 2 (solo mujeres), se observa que el p valor del test es de 0.7451, mismo que al ser mayor que 0.05 no se rechaza H_0, por lo cual, se concluye que los datos del modelo provienen de una población normal,lo cual es respaldado por el qqplot de los residuos de modelo, ya que se aprecia que la gran mayoría de puntos en el qqplot se ubican sobre la recta de normalidad. Adicionalmente, en cuanto al modelo 3 considerando ambos sexos de personas, se observa que el p valor de la prueba de normalidad es de 0.0387, el cual al ser menor que 0.05, por lo que se rechaza H_0, y se concluye que los datos de dicho modelo no provienen de una población normal, además esto se respalda por el qqplot donde se aprecia que los puntos del gráfico presentan un mayor alejamiento de la recta de normalidad que en los modelos previos, indicando que no hay normalidad de los datos, por último, en cuanto al modelo de la interacción entre sexo y estatura, se observa que el p valor del test de normalidad es igual a 0.0351, lo cual al se menor que 0.05 se rechaza H_0, por lo que los datos no provienen de una población normal, además de que en el qqplot se aprecia que existen alejamientos medianamente considerables de los puntos del gráfico respecto a la recta de normalidad, lo cual también sugiere que la normalidad no está presente entre los datos de éste último modelo.

2. Verificación de media 0

$$H_0: \mu = 0$$

$$H_1: \mu \neq 0$$

Modelo 1

```r
# test para comprobar que la media de residuos sea igual a 0 para modelo
1

t.test(modelo.hombres$residuals)
```

```
##
##  One Sample t-test
##
## data:  modelo.hombres$residuals
## t = 4.5495e-16, df = 219, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.4876507  0.4876507
## sample estimates:
##    mean of x
## 1.125698e-16
```

Modelo 2

```r
# test para comprobar que la media de residuos sea igual a 0 para modelo
2

t.test(modelo.mujeres$residuals)
```

```
##
##  One Sample t-test
##
## data:  modelo.mujeres$residuals
## t = -3.9979e-16, df = 219, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.881609  0.881609
## sample estimates:
##     mean of x
## -1.788342e-16
```

Modelo 3

```r
# test para comprobar que la media de residuos sea igual a 0 para modelo
3

t.test(modelo.HM$residuals)
```

```
##
##  One Sample t-test
##
```

```
## data:  modelo.HM$residuals
## t = 2.4085e-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
## 6.163788e-17
```

Modelo 4

```
# test para comprobar que la media de residuos sea igual a 0 para modelo
4

t.test(new.model$residuals)

##
##  One Sample t-test
##
## data:  new.model$residuals
## t = -8.5817e-16, df = 439, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.5017741  0.5017741
## sample estimates:
##     mean of x
## -2.190956e-16
```

En resumen, para todos los 4 modelos, el p valor del test es igual a 1, lo cual al ser mayor que 0.05, indica que en todos los casos, no se rechaza H_0 lo cual significa que efectivamente, la media de los residuos de cada modelo es igual a 0, cumpliendo de esa manera los 4 modelos con dicho supuesto.

3. Homocedasticidad

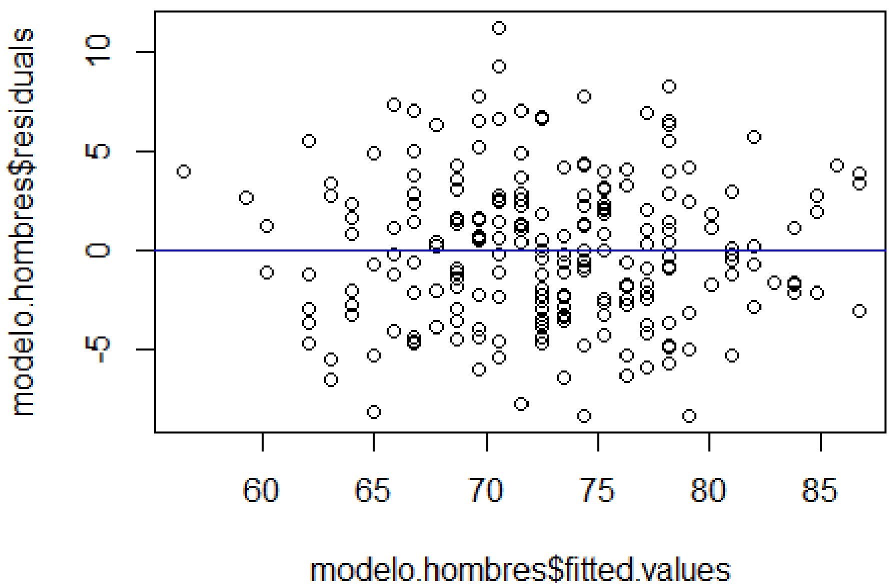
H_0: La varianza de los errores es constante (homocedasticidad) H_1: La varianza de los errores no es constante (heterocedasticidad)

```
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

Modelo 1

```
# Grafica de valores predichos vs residuos del modelo 1
```

```r
plot(modelo.hombres$fitted.values, modelo.hombres$residuals)
abline(h=0, col="blue")
```




```r
dwtest(modelo.hombres)
```

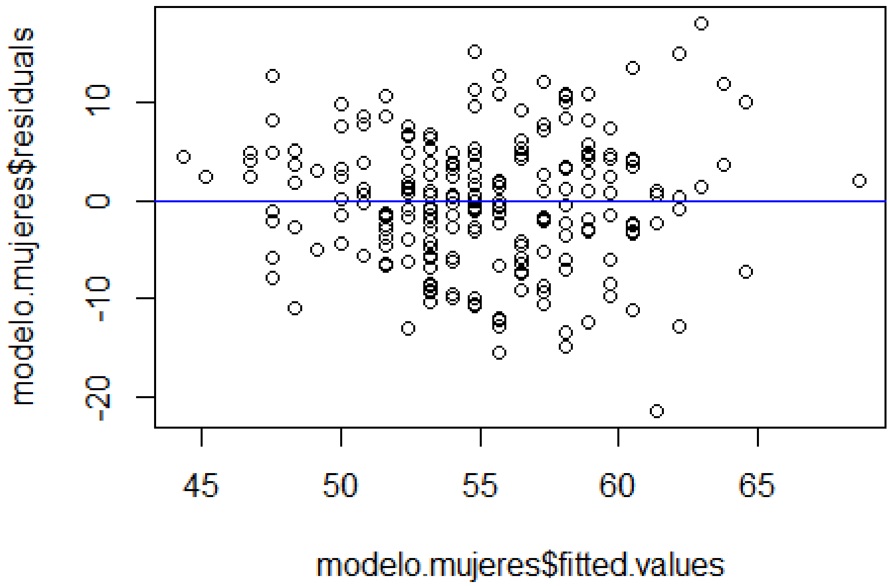
```
##
##  Durbin-Watson test
##
## data:  modelo.hombres
## DW = 2.0556, p-value = 0.6599
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
bgtest(modelo.hombres)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  modelo.hombres
## LM test = 0.20778, df = 1, p-value = 0.6485
```

Modelo 2

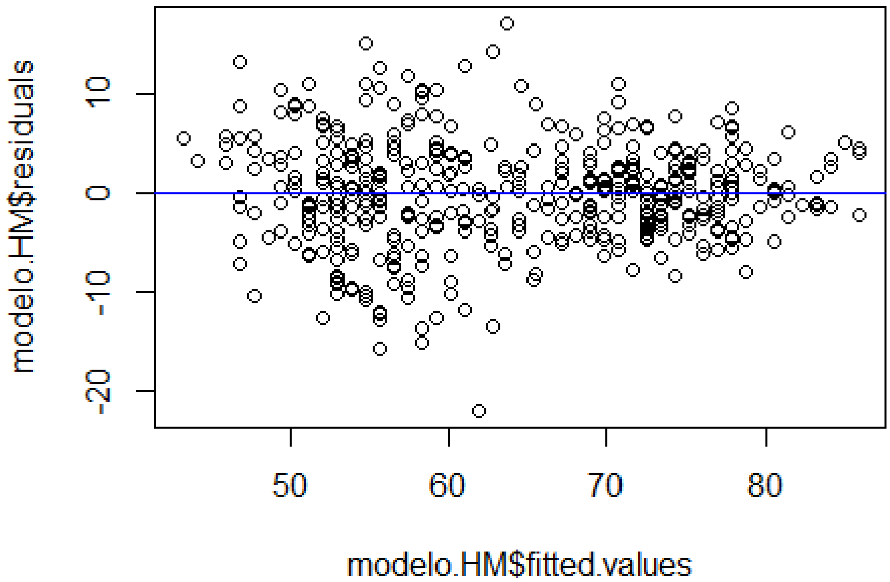
```r
# Grafica de valores predichos vs residuos del modelo 2

plot(modelo.mujeres$fitted.values, modelo.mujeres$residuals)
abline(h=0, col="blue")
```




```
dwtest(modelo.mujeres)

##
##  Durbin-Watson test
##
## data:  modelo.mujeres
## DW = 1.8062, p-value = 0.07532
## alternative hypothesis: true autocorrelation is greater than 0

bgtest(modelo.mujeres)

##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  modelo.mujeres
## LM test = 1.4655, df = 1, p-value = 0.2261
```

Modelo 3

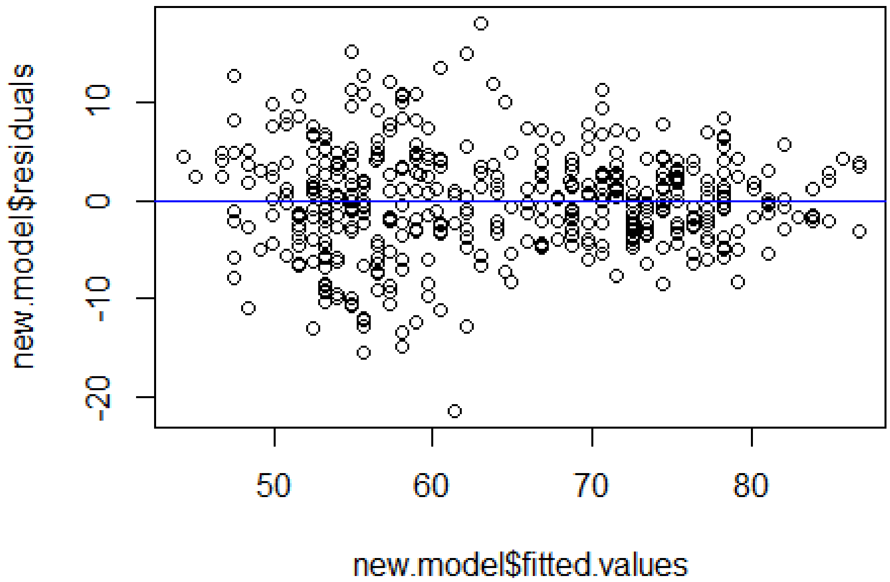
```
# Grafica de valores predichos vs residuos del modelo 3

plot(modelo.HM$fitted.values, modelo.HM$residuals)
abline(h=0, col="blue")
```




```
dwtest(modelo.HM)

##
##  Durbin-Watson test
##
## data:  modelo.HM
## DW = 1.8663, p-value = 0.07325
## alternative hypothesis: true autocorrelation is greater than 0

bgtest(modelo.HM)

##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  modelo.HM
## LM test = 1.3595, df = 1, p-value = 0.2436
```

Modelo 4

```
# Grafica de valores predichos vs residuos de la interacción estatura-
sexo

plot(new.model$fitted.values, new.model$residuals)
abline(h=0, col="blue")
```




```
dwtest(new.model)

## 
##  Durbin-Watson test
## 
## data:  new.model
## DW = 1.8646, p-value = 0.07113
## alternative hypothesis: true autocorrelation is greater than 0

bgtest(new.model)

## 
##  Breusch-Godfrey test for serial correlation of order up to 1
## 
## data:  new.model
## LM test = 1.3453, df = 1, p-value = 0.2461
```

Finalmente, en cuanto al modelo 1, el p valor de los test de normalidad de Durbin Watson y Breusch-Godfrey son 0.65 y 0.64 respectivamente, por lo cual, en caso del modelo 1, para el caso de ambos tests, no se rechaza H_0, por lo que se puede concluir que la varianza de los residuos es constante, además, para el caso del modelo 2, el p valor de las 2 pruebas es de 0.07 y 0.22, por lo que en ambos casos no se rechaza H_0, por lo que se concluye que la varianza de los residuos del modelo es constante. Además, en el caso del modelo 3,el p valor de ambas pruebas es de 0.07 y 0.24, por lo cual en ambos tests no se rechaza H_0, motivo por el cual se concluye que la varianza de los residuos en este modelo también es constante. Finalmente, para el caso del modelo 4, el p valor de los tests es de 0.07 y 0.24 respectivamente, por lo que al ser ambos valores mayores a 0.05, no se rechaza H_0, lo cual indica que los residuos de este último modelo también presentan varianza constante, lo cual para el caso de los 4 modelos es respaldado por el hecho de que en el gráfico se predicciones vs residuos se aprecia que los puntos de los gráficos están dispersos de forma mayormente homogénea alrededor de la media de los residuos.

4. Independencia

H_0: Los errores no están correlacionados H_1: Los errores están correlacionados

Modelo 1

```
# Test de Breusch Pagan para independencia de residuos

bptest(modelo.hombres)

##
##  studentized Breusch-Pagan test
##
## data:  modelo.hombres
## BP = 0.93324, df = 1, p-value = 0.334

# Test de Goldfend Quandt para independencia de residuos

gqtest(modelo.hombres)

##
##  Goldfeld-Quandt test
##
## data:  modelo.hombres
## GQ = 0.84148, df1 = 108, df2 = 108, p-value = 0.8144
## alternative hypothesis: variance increases from segment 1 to 2
```

Modelo 2

```
# Test de Breusch Pagan para independencia de residuos

bptest(modelo.mujeres)

##
##  studentized Breusch-Pagan test
##
## data:  modelo.mujeres
## BP = 8.4976, df = 1, p-value = 0.003556

# Test de Goldfend Quandt para independencia de residuos

gqtest(modelo.mujeres)

##
##  Goldfeld-Quandt test
##
## data:  modelo.mujeres
```

```
## GQ = 1.4265, df1 = 108, df2 = 108, p-value = 0.03313
## alternative hypothesis: variance increases from segment 1 to 2
```

Modelo 3

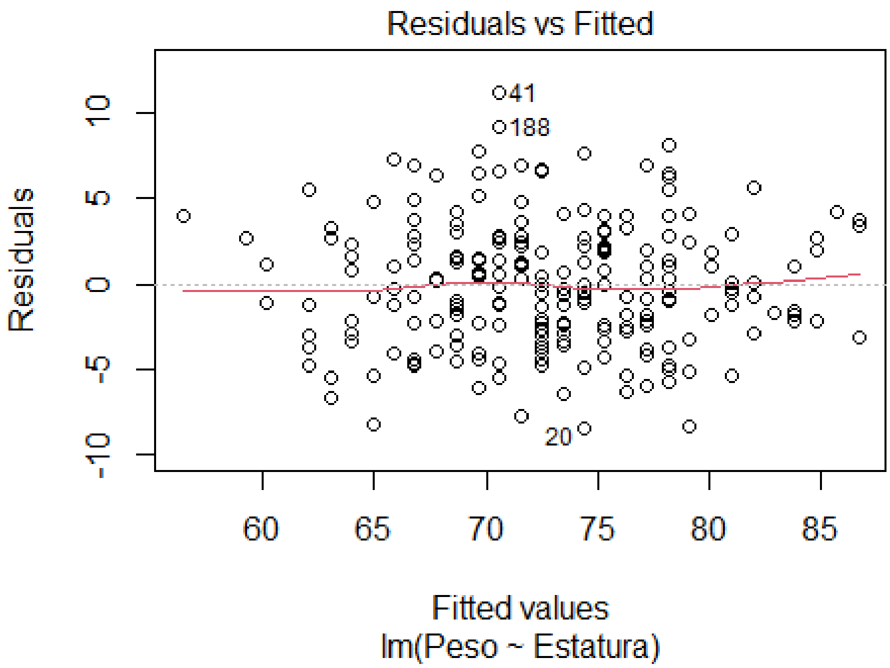
```
# Test de Breusch Pagan para independencia de residuos

bptest(modelo.HM)

##
##  studentized Breusch-Pagan test
##
## data:  modelo.HM
## BP = 48.202, df = 2, p-value = 3.413e-11

# Test de Goldfend Quandt para independencia de residuos

gqtest(modelo.HM)

##
##  Goldfeld-Quandt test
##
## data:  modelo.HM
## GQ = 3.2684, df1 = 217, df2 = 217, p-value < 2.2e-16
## alternative hypothesis: variance increases from segment 1 to 2
```

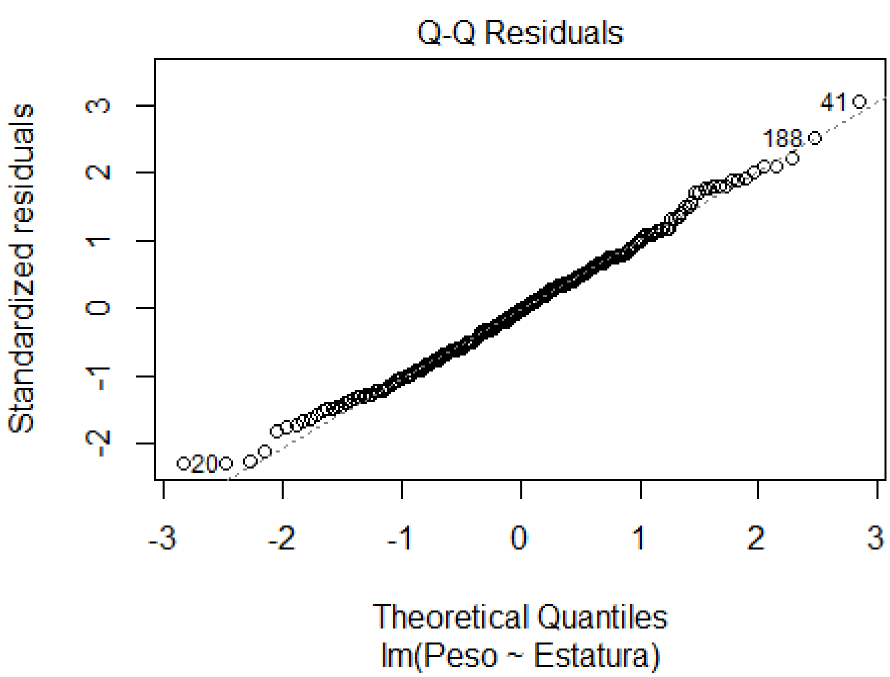
Modelo 4

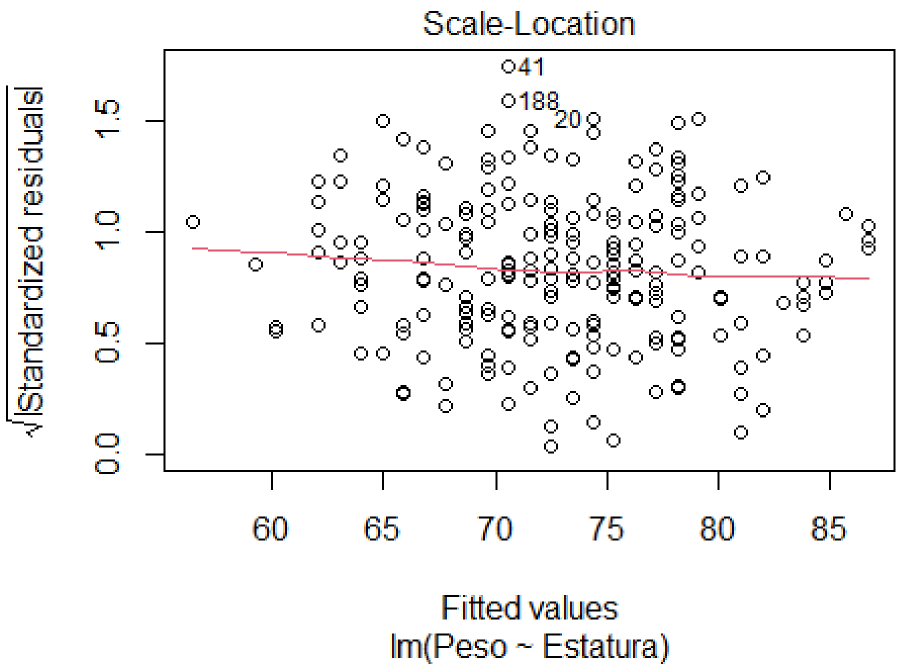
```
# Test de Breusch Pagan para independencia de residuos

bptest(new.model)

##
##  studentized Breusch-Pagan test
##
## data:  new.model
## BP = 59.211, df = 3, p-value = 8.667e-13

# Test de Goldfend Quandt para independencia de residuos

gqtest(new.model)

##
##  Goldfeld-Quandt test
##
## data:  new.model
## GQ = 3.2684, df1 = 216, df2 = 216, p-value < 2.2e-16
## alternative hypothesis: variance increases from segment 1 to 2
```

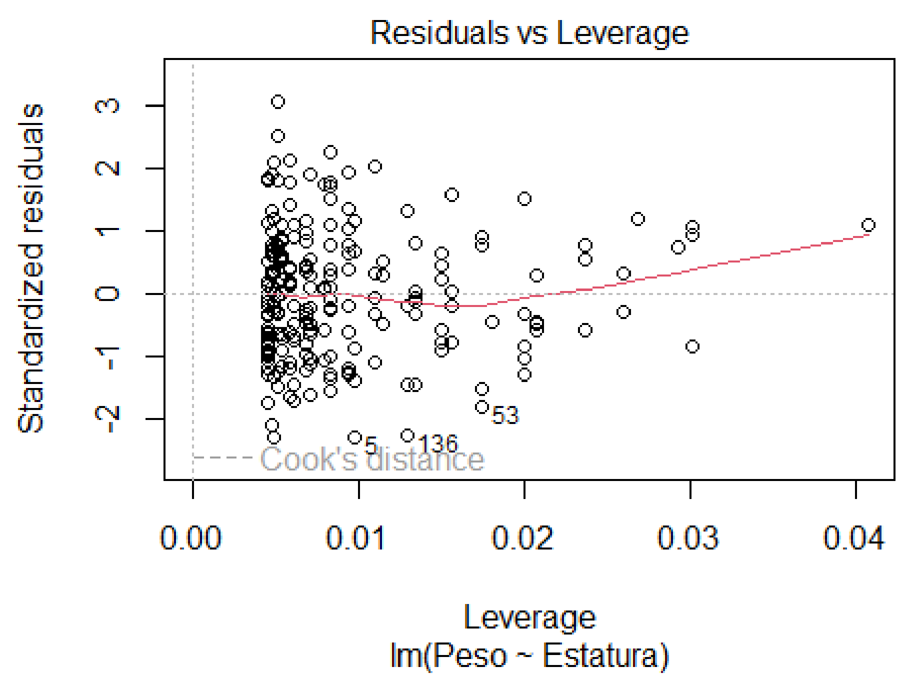
En los tests de independencia de Breusch-Pagan y Goldfeld-Quandt realizados anteriormente para cada uno de los modelos, se aprecia que para el caso del modelo 1, los valores p de los tests son de 0.33 y 0.81, los cuales al ser superiores a 0.03, no se rechaza H_0, por lo que en el caso del modelo 1, se concluye que los residuos del modelo no están correlacionados, es decir que son independientes, además en cuanto al modelo 2, los p valores de los tests de independencia son de 0.003 y 0.03, por lo que en el caso del primer test se rechaza H_0 mientras que según el segundo test no se rechaza H_0, sin embargo, observando las gráficas de predicciones vs residuos realizadas previamente, se observa que para el caso del modelo 2 los puntos no siguen alguna tendencia o patrón en específico, por lo que se concluye que los residuos del modelo 2 son independientes. Además, para el modelo 3 los p valores de los tests son 3.4e-11 y el segundo inferior a 2.2e-16, por lo que en ambos casos se rechaza H_0, indicando que los residuos del modelo 3 sí están correlacionados, en otras palabras, que no son independientes y por último en cuanto al modelo 4, los p valores de los tests son 8.667e-13 y el segundo inferior a 2.2e-16, por lo cual, en el caso del modelo 4 se concluye que los errores del modelo están correlacionados y por tanto no son independientes.

```
# Graficar modelo 1

plot(modelo.hombres)
```

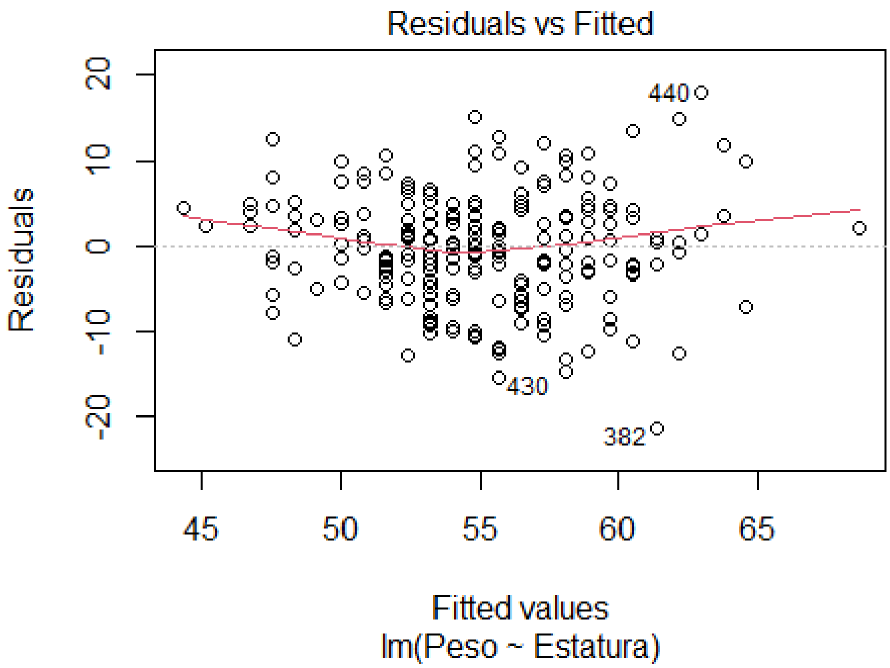

Residuals vs Fitted
Residuals
Fitted values
lm(Peso ~ Estatura)
41
188
20

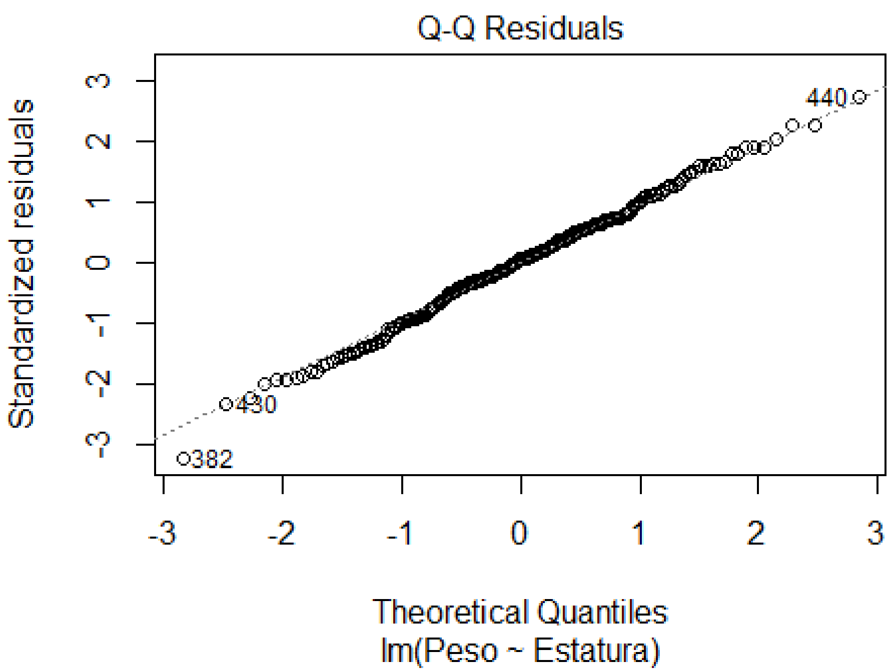

Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Peso ~ Estatura)
41
188
20

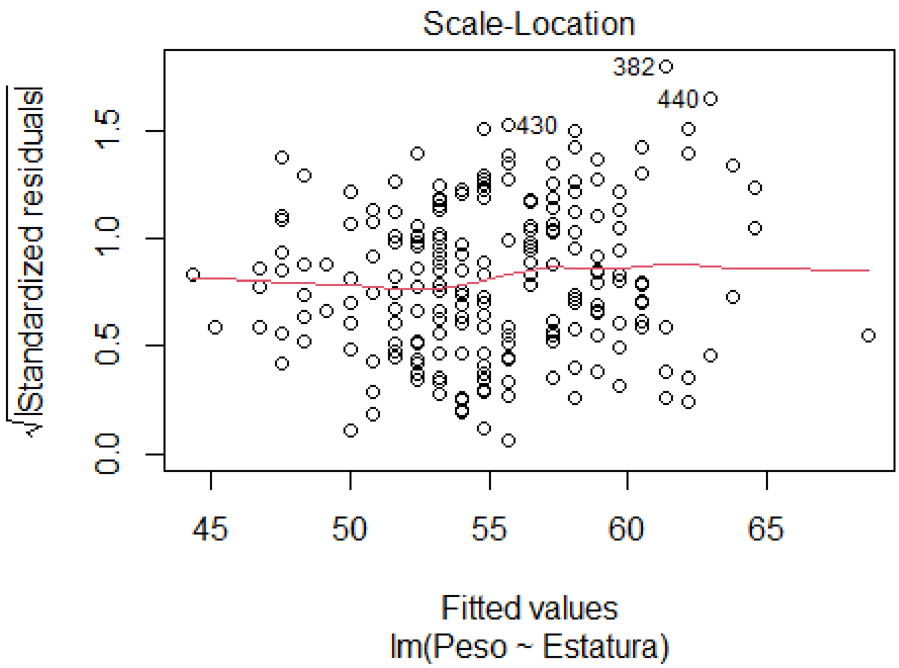


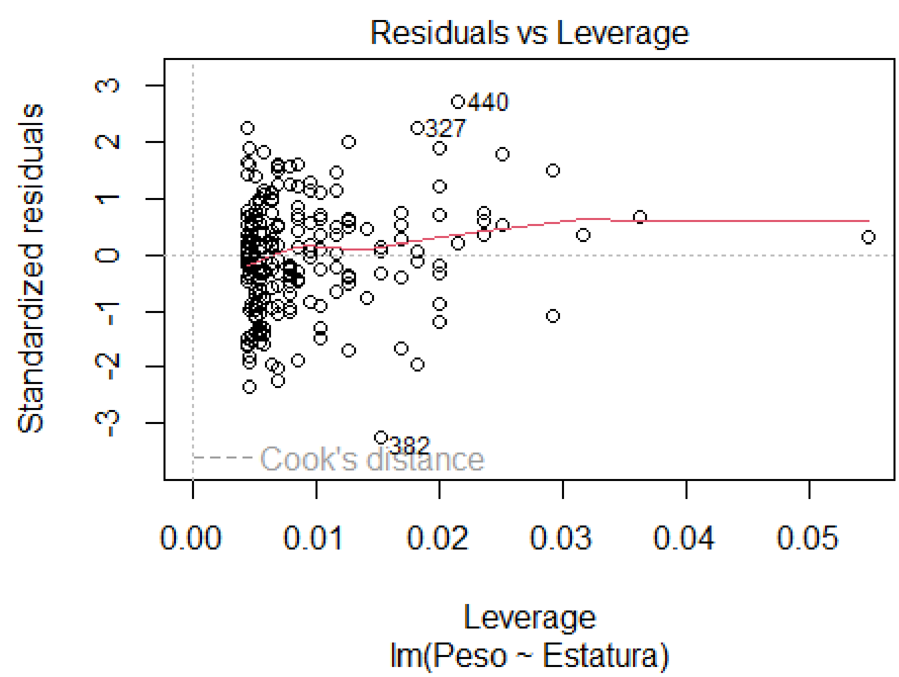



```r
# Graficar modelo 2

plot(modelo.mujeres)
```

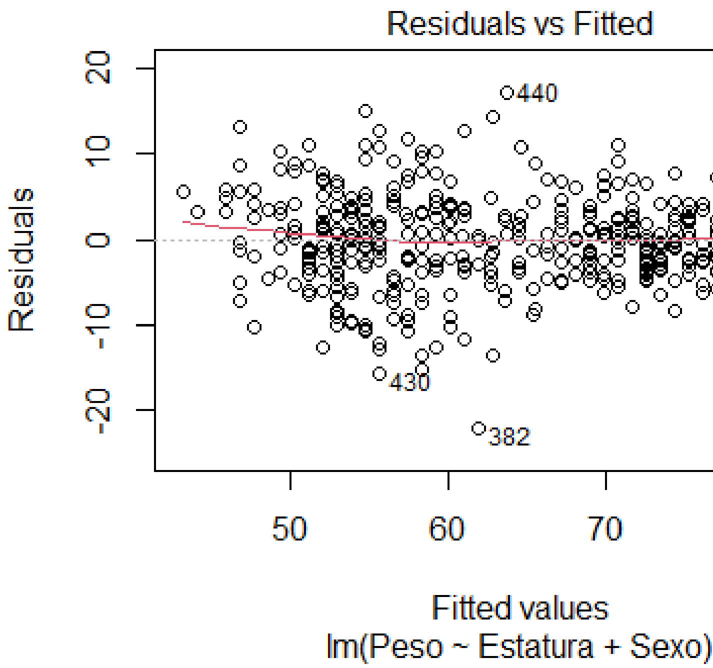

Residuals vs Fitted
Residuals
Fitted values
lm(Peso ~ Estatura)
440
430
382


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Peso ~ Estatura)
440
430
382

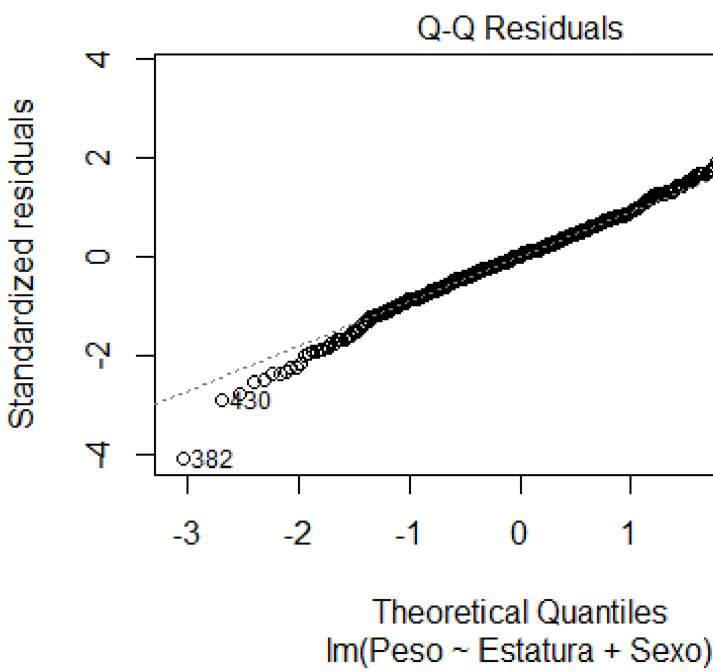




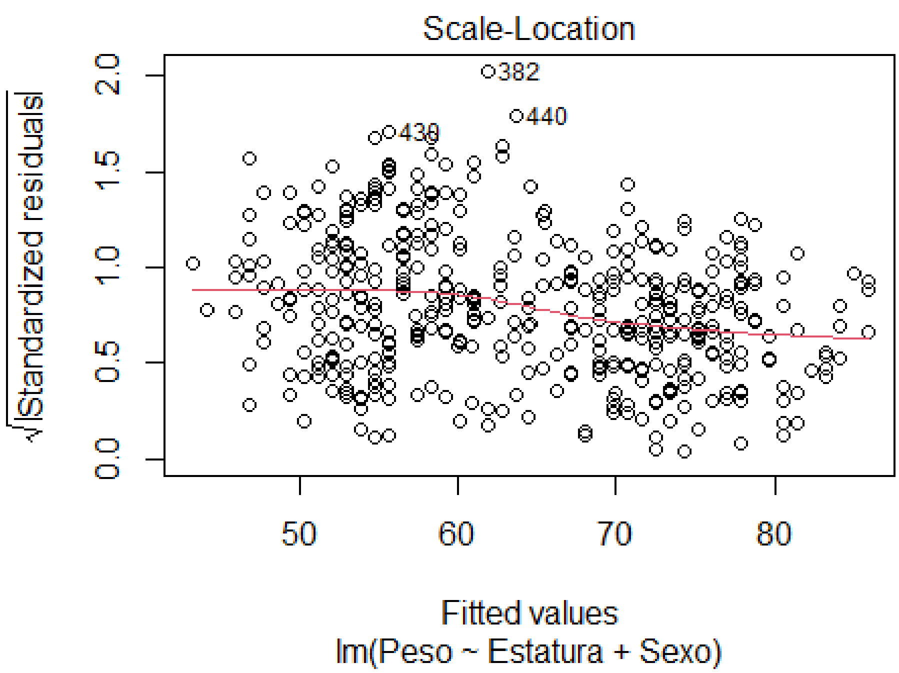

```
# Graficar modelo 3

plot(modelo.HM)
```

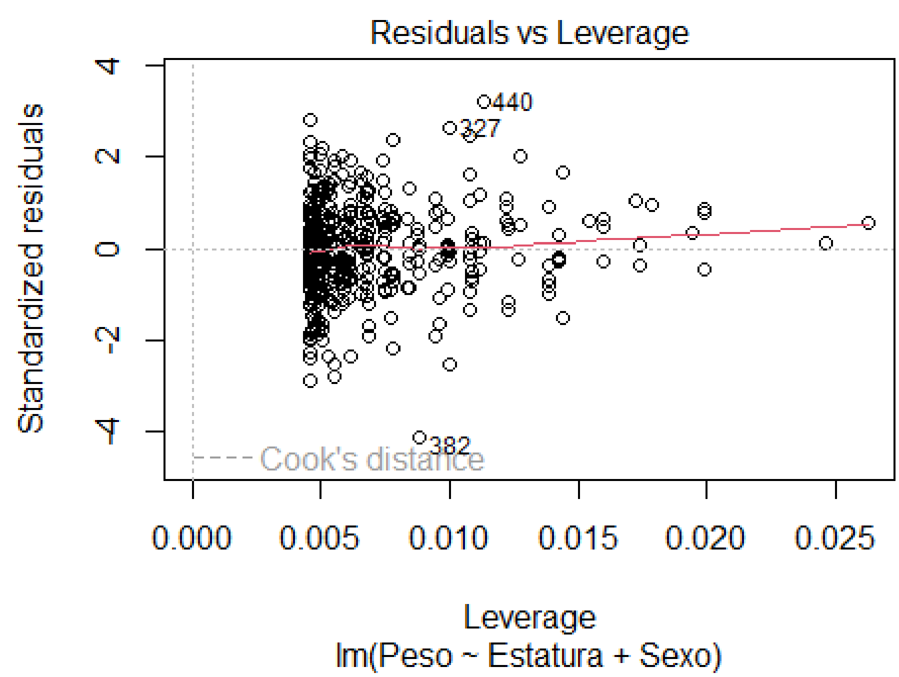


Residuals vs Fitted
Residuals
20
10
0
-10
-20
440
430
382
50
60
70
80
Fitted values
lm(Peso ~ Estatura + Sexo)

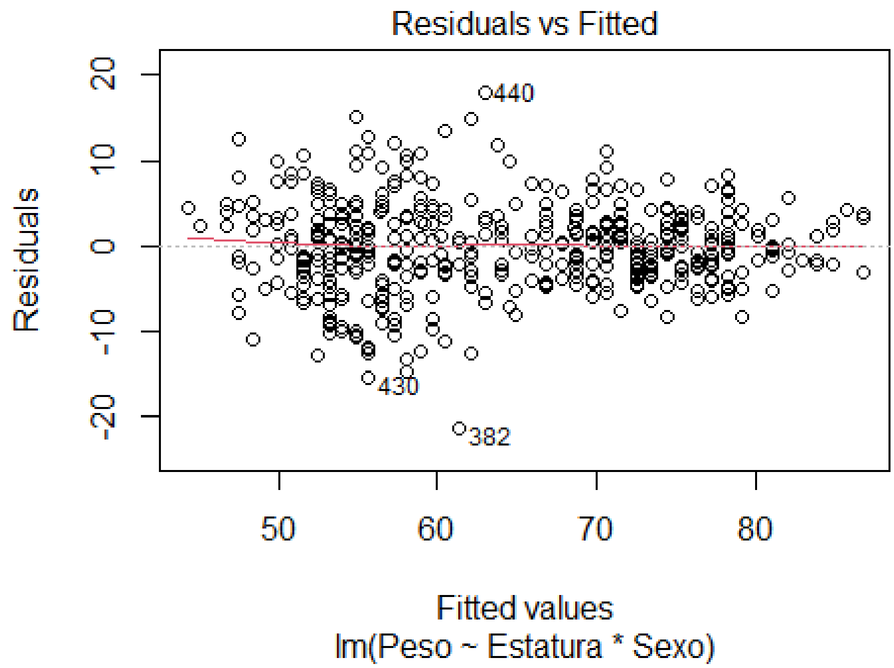


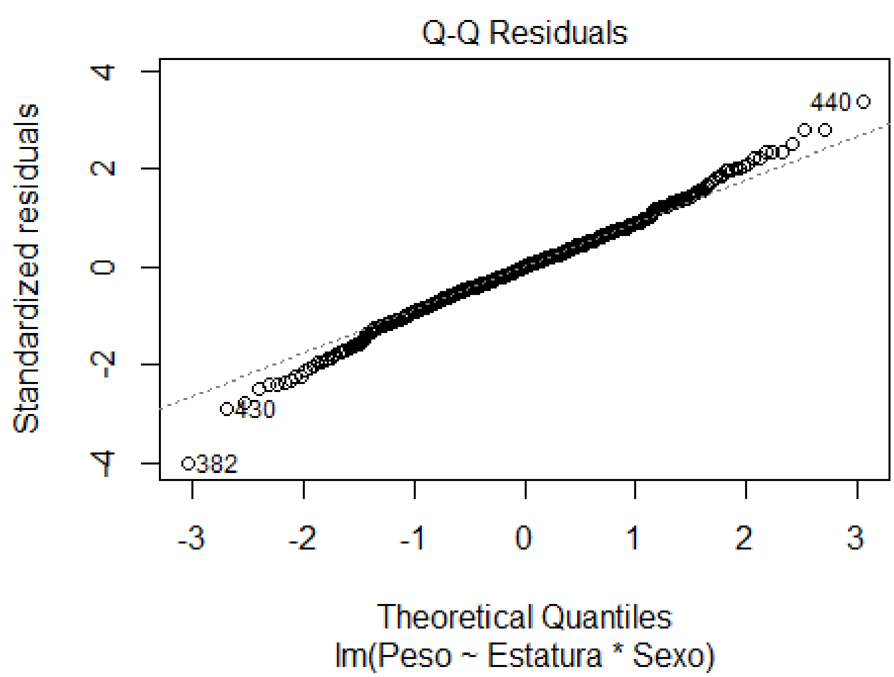
Q-Q Residuals
Standardized residuals
4
2
0
-2
-4
440
430
382
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Peso ~ Estatura + Sexo)

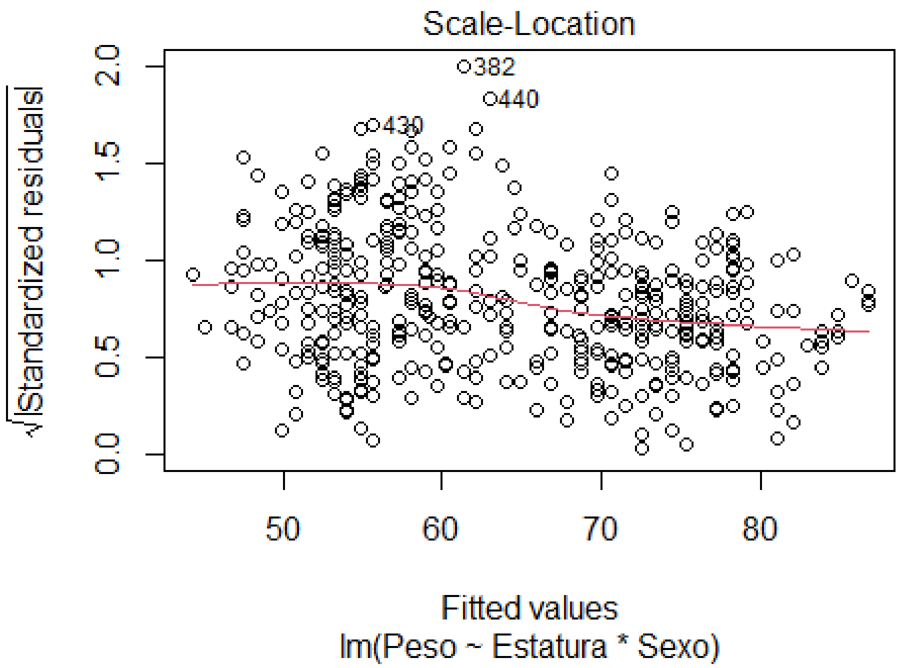


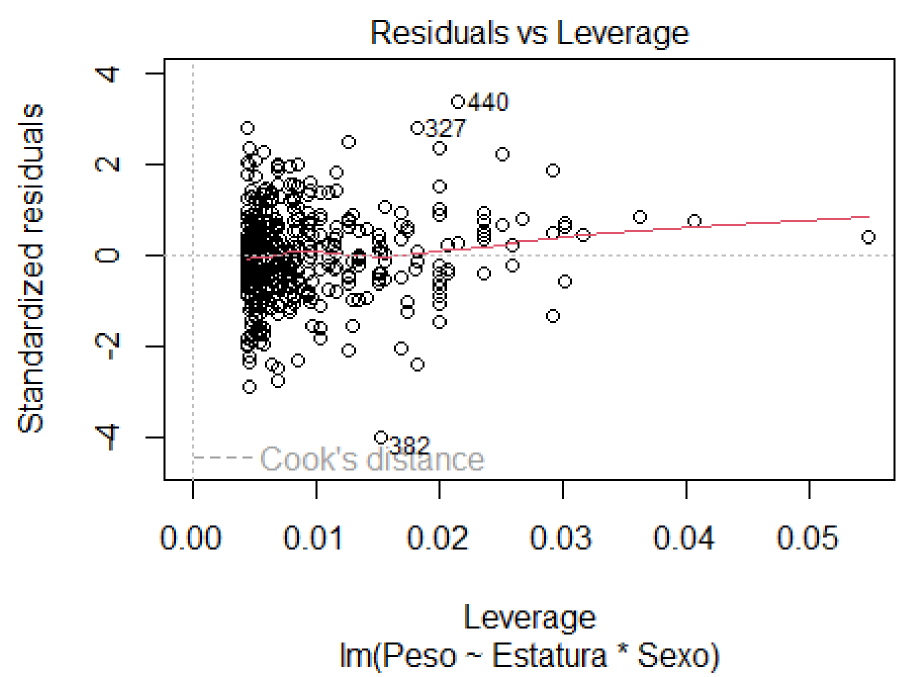



```
# Graficar modelo 4

plot(new.model)
```


Residuals vs Fitted
Residuals
440
430
382
50
60
70
80
Fitted values
lm(Peso ~ Estatura * Sexo)


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





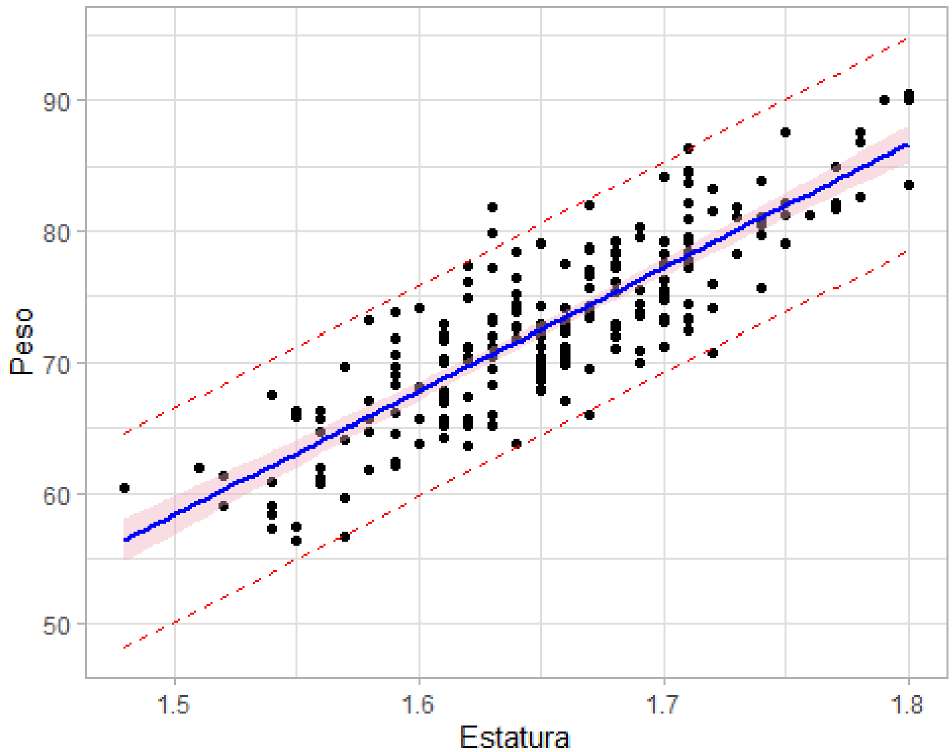
En general, en las gráficas de los modelos se observa que dichas gráficas son bastante similares a las que ya se habían analizado previamente, en el sentido de que los puntos de los gráficos se dispersan de manera mayormente homogénea a lo largo de los gráficos, además de que también los puntos representados en el qqplot se encuentran en su gran mayoría ubicados sobre la recta que representa la normalidad de los datos, por lo cual es posible afirmar que éstos gráficos derivados de los modelos no cambian las conclusiones obtenidas anteriormente con respecto a normalidad, media 0, independencia y homocedasticidad de los residuos para cada uno de los modelos.

Conclusión final:

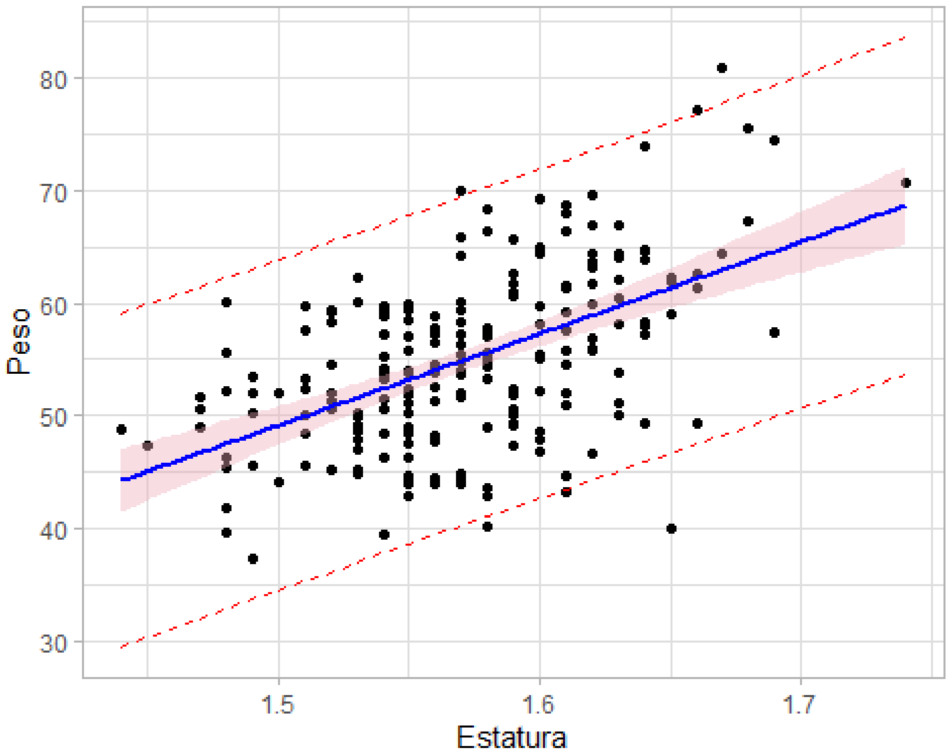
Finalmente se concluye que el mejor modelo de regresión lineal para representar los datos en cuestión es el referente a la relación existente entre el peso y la estatura de las personas de ambos sexos, tanto hombres como mujeres, dado que es aquel modelo que tiene el mayor valor p referente a normalidad de sus residuos, indicando que es el modelo que tiene la mayor probabilidad de todas de cumplir con la normalidad entre sus residuos, además este modelo también cumple con otros supuestos de validación, tales como homocedasticidad, media 0 e independencia de sus residuos, mientras que los otros 3 modelos incumplen con al menos uno de los supuestos de validación de los modelos de regresión lineal, por lo cual, el modelo que asegurará obtener las predicciones más precisas y confiables para el peso de las personas a partir de su estatura es el modelo que contempla tanto a hombres como a mujeres (modelo 3).

Intervalos de confianza

```
# libreria para realizar graficos personalizados

library(ggplot2)

# Intervalos de prediccion para hombres
suppressWarnings({
Ip=predict(object=modelo.hombres,interval="prediction",level=0.97)
datos=cbind(datos_H, Ip)

ggplot(datos,aes(x=Estatura,y=Peso))+
geom_point()+
geom_line(aes(y=lwr), color="red", linetype="dashed")+
geom_line(aes(y=upr), color="red", linetype="dashed")+
geom_smooth(method=lm, formula= y~x, se=TRUE, level=0.97, col="blue",
fill="pink2")+
theme_light()})
```




```
#Intervalos de prediccion para mujeres

suppressWarnings({

Ip=predict(object=modelo.mujeres,interval="prediction",level=0.97)
datos=cbind(datos_M, Ip)

ggplot(datos,aes(x=Estatura,y=Peso))+
geom_point()+
geom_line(aes(y=lwr), color="red", linetype="dashed")+
geom_line(aes(y=upr), color="red", linetype="dashed")+
geom_smooth(method=lm, formula= y~x, se=TRUE, level=0.97, col="blue",
fill="pink2")+
theme_light()})
```



Por último, en cada una de las gráficas de los intervalos de predicción para ambos modelos, se aprecia que las línea rojas punteadas hacen referencia al intervalo dentro del cual se espera que se ubique la siguiente observación con un nivel de confianza del 97%, mientras qie al mismo tiempo, la recta azul en ambos casos simboliza el modelo de regresión obtenido para los datos tanto de hombres como de mujeres, además dicha recta azul hace referencia al lugar donde el modelo predice que estará el peso promedio de las personas de la próxima observación que se lleve a cabo, mientras que el área en color rojo representa el intervalo de confianza para cada uno de los modelos de regresión realizados, dicho de otra forma el lugar donde se espera que ubique el valor real de la próxima observación que se realice con un nivel de confianza del 97%.