act12_A01742161

Rogelio Lizárraga

2024-09-04

La recta de mejor ajuste (Tercera entrega)

```
library(ggplot2)
```

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.

```
M =  read.csv('Estatura-peso_HyM.csv')
MM = subset(M,M$Sexo=="M")
MH = subset(M,M$Sexo=="H")
M1=data.frame(MH$Estatura,MH$Peso,MM$Estatura,MM$Peso)
cor(M1)
```

```
##              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
```

```
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
```

Observamos que la estatura y el peso están fuertemente correlacionados para los hombres, pero no tanto para las mujeres. Además, vemos que la media y la mediana están muy cercanas en la estatura de los hombres y mujeres, y en el peso de los hombres, pero estas se encuentran más alejadas entre sí para las mujeres. También, observamos que la desviación estándar es mayor para el peso de las mujeres que para el peso de los hombres.

3. Encuentra la ecuación de regresión de mejor ajuste:

Probaremos primero el modelo de regresión lineal para la base de datos de hombres: ## a. Realiza la regresión entre las variables involucradas

```
Modelo1H  = lm(Peso~Estatura, MH)
Modelo1H
```

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

Para esta tenemos la ecuación

$$Peso = -83.68 + 94.66 Estatura$$

b. Verifica el modelo:

- $H_0 : \beta_i = 0$
- $H_1 : \beta_i \neq 0$

```
summary(Modelo1H)
```

```
##
## Call:
## lm(formula = Peso ~ Estatura, data = MH)
##
## 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
```

Verifica la significancia del modelo con un alfa de 0.03. Verifica la significancia de $\hat{\beta}_i$ con un alfa de 0.03. Verifica el porcentaje de variación explicada por el modelo

Observamos que el valor p es < 0.03 para β_0 y β_1, por lo que se rechaza H_0 y llegamos a la conclusión que los coeficientes del intercepto y la estatura sí son estadísticamente significativos. De la misma manera, el modelo tiene una significancia menor a 0.03, por lo que el modelo es estadísticamente significativo. Finalmente, tenemos un coeficiente de determinación del 0.7171, por lo que el modelo (la estatura de los hombres) explica el 71.71% de la varianza en el peso de los hombres.

Ahora, probaremos el modelo de regresión lineal para la base de datos de mujeres: ## a. Realiza la regresión entre las variables involucradas

```
Modelo1M  = lm(Peso~Estatura, MM)
Modelo1M
```

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

Para esta tenemos la ecuación

$$Peso = -72.56 + 81.15 Estatura$$

b. Verifica el modelo:

- $H_0 : \beta_i = 0$
- $H_1 : \beta_i \neq 0$

```
summary(Modelo1M)
```

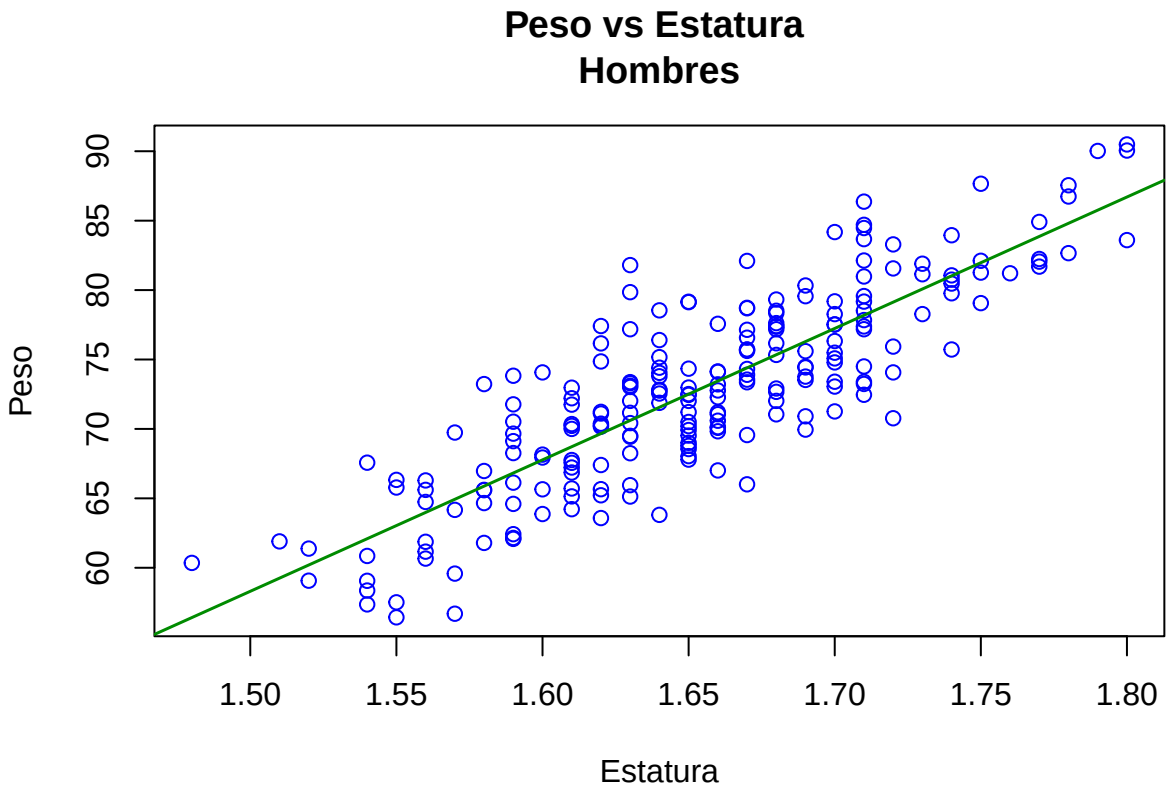
```
##
## Call:
## lm(formula = Peso ~ Estatura, data = MM)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3256  -4.1942   0.4004   4.2724  17.9114
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -72.560     14.041  -5.168 5.34e-07 ***
## 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
```

Verifica la significancia del modelo con un alfa de 0.03. Verifica la significancia de $\hat{\beta}_i$ con un alfa de 0.03. Verifica el porcentaje de variación explicada por el modelo

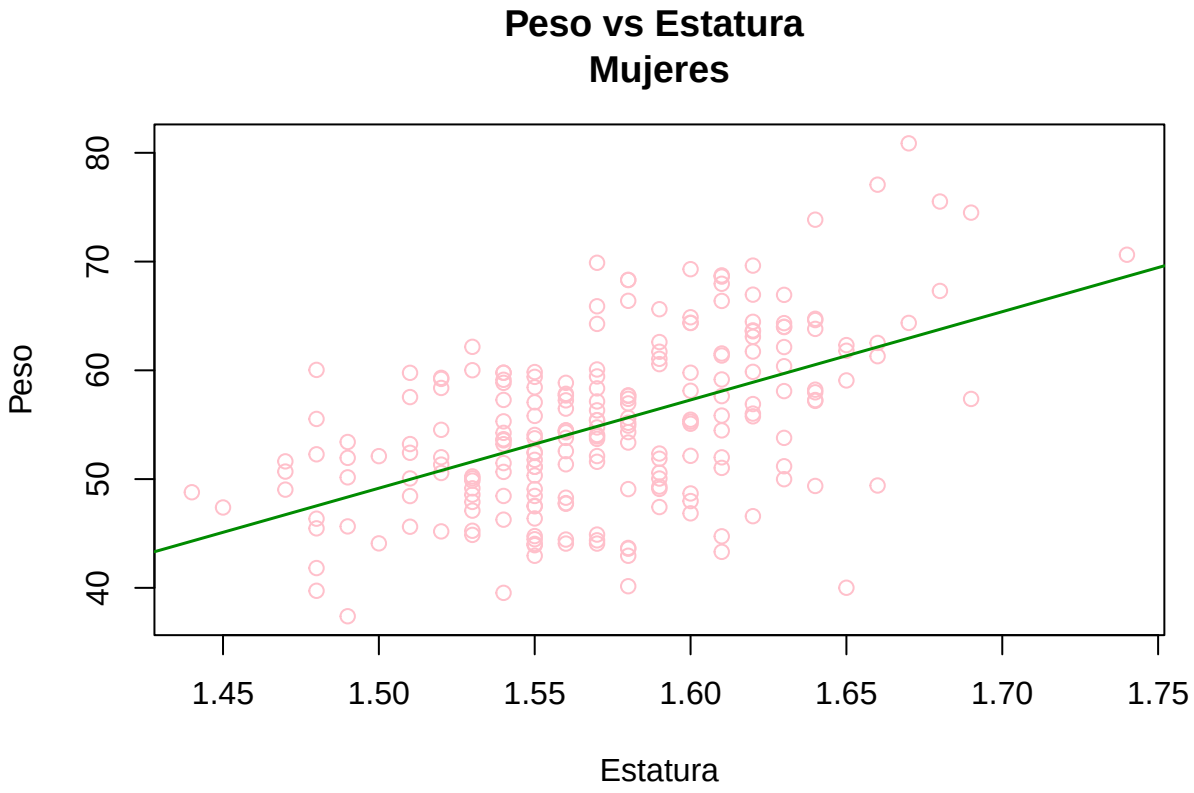
Observamos que el valor p es < 0.03 para β_0 y β_1, por lo que se rechaza H_0 y llegamos a la conclusión que los coeficientes del intercepto y la estatura sí son estadísticamente significativos. De la misma manera, el modelo tiene una significancia menor a 0.03, por lo que el modelo es estadísticamente significativo. Es decir, sí hay una asociación entre las variable dependiente e independiente (peso y estatura). Finalmente, tenemos un coeficiente de determinación del 0.2718, por lo que el modelo (la estatura de las mujeres) explica el 27.18% de la varianza en el peso de las mujeres, lo cual es una variación explicada muy baja.

4. Dibuja el diagrama de dispersión de los datos y las rectas.

```
plot(MH$Estatura, MH$Peso, col ='blue', main = 'Peso vs Estatura \n Hombres ', ylab = 'Peso', xlab = 'Es
abline(Modelo1H, col = 'green4', lwd = 1.5, pch = 19)
```




```
plot(MM$Estatura, MM$Peso, col ='pink', main = 'Peso vs Estatura \n Mujeres ', ylab = 'Peso', xlab = 'Es
abline(Modelo1M, col = 'green4', lwd = 1.5, pch = 19)
```



Posteriormente, haremos el modelo del peso y estatura de hombres y mujeres en conjunto.

```
Modelo2 = lm(Peso~ Estatura + Sexo, M)
Modelo2
```

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

En este modelo tenemos la ecuación

$$Peso = -74.75 + 89.26 Estatura - 10.56 SexoM$$

, donde $SexoM$ es una variable dummy que indica que: si es mujer, $SexoM = 1$ y $SexoM = 0$ si es otro caso.

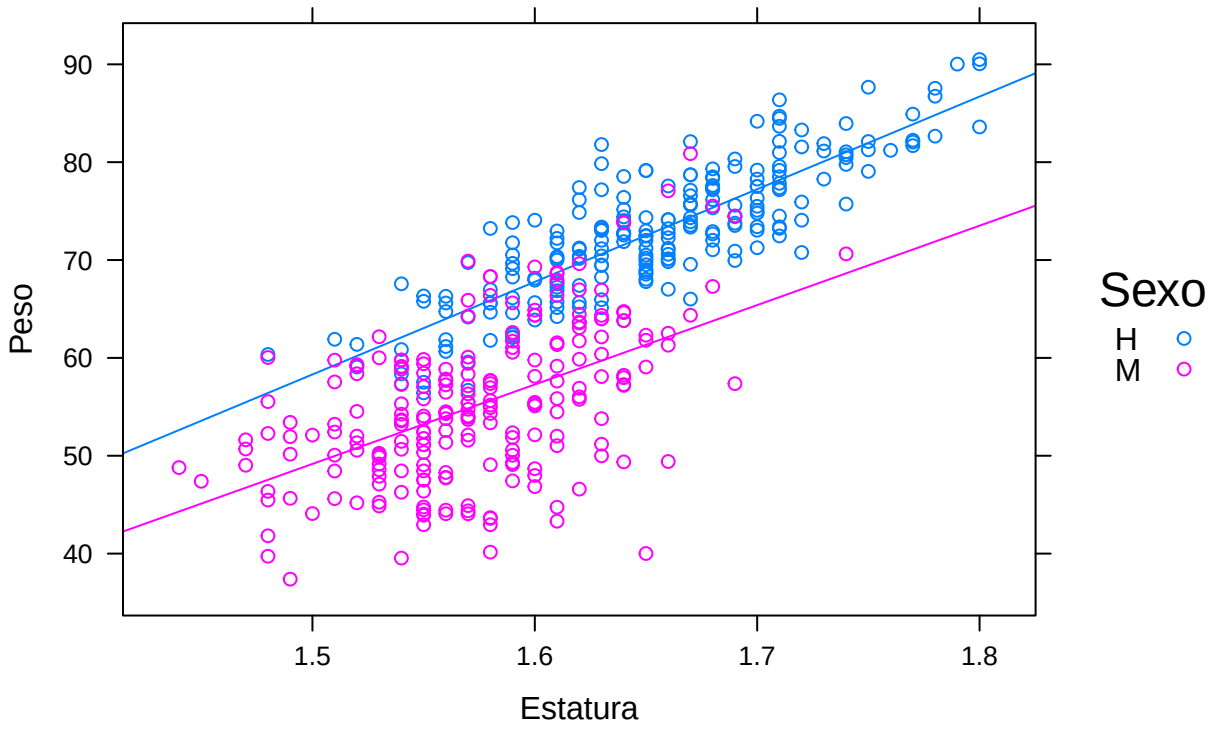
```
summary(Modelo2)
```

```
##
## Call:
## lm(formula = Peso ~ Estatura + Sexo, data = M)
```

```
##
## 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
```

Observamos que el valor p es < 0.03 para β_0, β_1 y $beta_2$, por lo que se rechaza H_0 y llegamos a la conclusión que los coeficientes del intercepto, la estatura y el sexo sí son estadísticamente significativos. De la misma manera, el modelo tiene una significancia menor a 0.03, por lo que el modelo es estadísticamente significativo. Es decir, sí hay una asociación entre las variable dependientes e independientes (peso, contra estatura y sexo). Finalmente, tenemos un coeficiente de determinación del 0.7827, por lo que el modelo (estatura + sexo) explica el 78.27% de la varianza en el peso de las personas, lo cual es una explicación buena.

```r
library(lattice)
modelo <- lm(Peso ~ Estatura + Sexo, data = M)
xyplot(Peso ~ Estatura, data = M, groups = Sexo,
       auto.key = list(space = "right", title = "Sexo"),
       type = c("p", "r"),
       xlab = "Estatura", ylab = "Peso",
       main = "Modelo sin interacción: Estatura y Sexo contra peso")
```



Continuación

Anteriormente, realizamos modelos para explicar la variación del peso teniendo en cuenta la estatura, para hombres y mujeres por separado, y analizando todo el conjunto, utilizando la estatura y el sexo como variables independientes sin interacción. Ahora, crearemos un modelo de regresión lineal con la estatura y sexo CON interacción:

```
Modelo3 = lm(Peso ~ Estatura * Sexo, M)
Modelo3
```

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

En este modelo tenemos la ecuación

$$Peso = -83.68 + 94.66 Estatura + 11.12 SexoM - 13.51 Estatura : SexoM$$

, donde $SexoM$ es una variable dummy que indica que: si es mujer, $SexoM = 1$ y $SexoM = 0$ si es otro caso. Por otro lado Estatura:SexoM que el impacto de Estatura en el Peso es 13.51 unidades menor para las mujeres en comparación con los hombres.

```r
summary(Modelo3)
```

```
##
## Call:
## lm(formula = Peso ~ Estatura * Sexo, data = M)
##
## 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
```

Observamos que el valor p es < 0.03 para β_0 y β_1, por lo que se rechaza H_0 para estos coeficientes y llegamos a la conclusión que los coeficientes del intercepto y la estatura sí son estadísticamente significativos. Por otro lado, el valor p es > 0.03 para β_2 y β_3, por lo que no se rechaza H_0 para estos coeficientes y concluimos que el coeficiente del sexo y el coeficiente de la interacción entre la estatura y el sexo no son estadísticamente significativos. De la misma manera, el modelo tiene una significancia menor a 0.03, por lo que el modelo es estadísticamente significativo. Es decir, sí hay una asociación entre las variable dependientes e independientes (peso, contra las variables independientes del modelo). Finalmente, tenemos un coeficiente de determinación del 0.7832, por lo que el modelo (estatura + sexo) explica el 78.32% de la varianza en el peso de las personas, lo cual es una explicación buena.

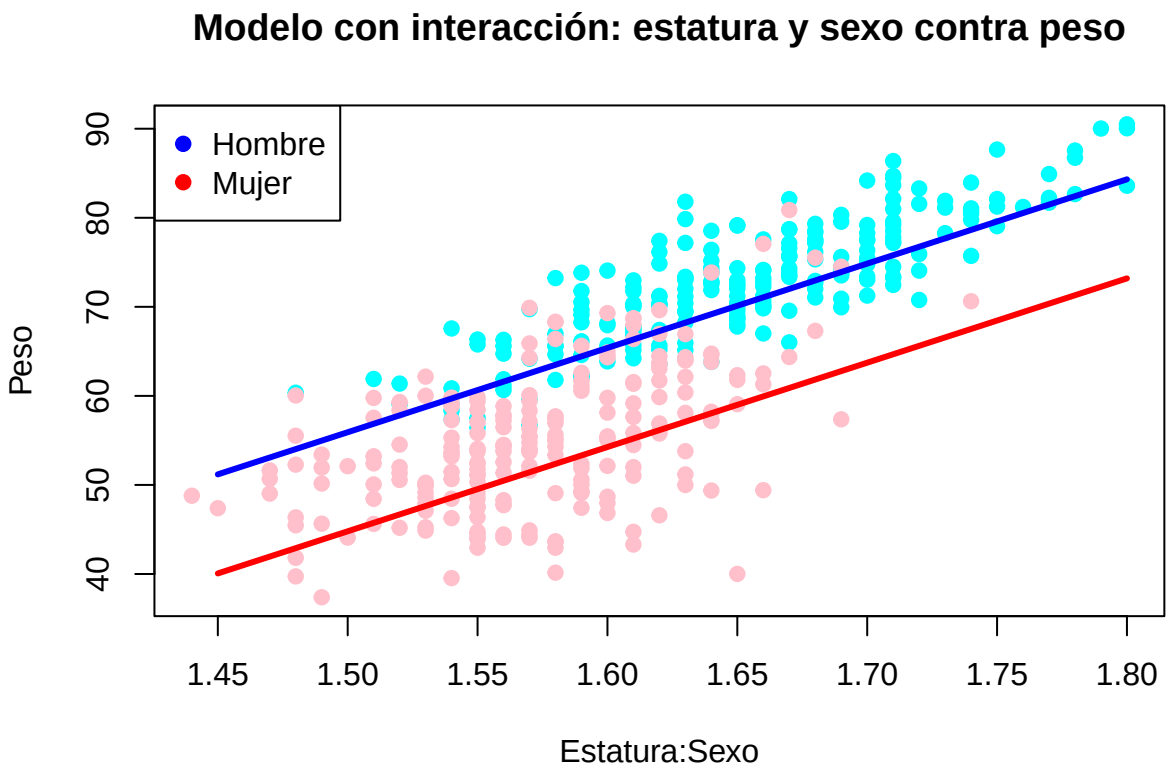
Sin embargo, al tener coeficientes que no son estadísticamente significativos (la interacción entre la estatura y el sexo, y el sexo), este modelo no es el más adeucado.

```r
beta_0=Modelo3$coefficients[1]
beta_1=Modelo3$coefficients[2]
beta_2=Modelo3$coefficients[3]
beta_3 = Modelo3$coefficients[4]

Ym=function(x){beta_0+beta_2+beta_1*x+beta_3}
Yh=function(x){beta_0+beta_1*x+beta_3}

colores=c("cyan","pink")
plot(M$Estatura,M$Peso,col=colores[factor(M$Sexo)],pch=19,ylab="Peso",xlab="Estatura:Sexo",main="Modelo
x=seq(1.45,1.80,0.01)
lines(x,Ym(x),col="blue",lwd=3)
lines(x,Yh(x),col="red",lwd=3)

legend("topleft", legend=c("Hombre","Mujer"), pch=19, col=c("blue","red"))
```



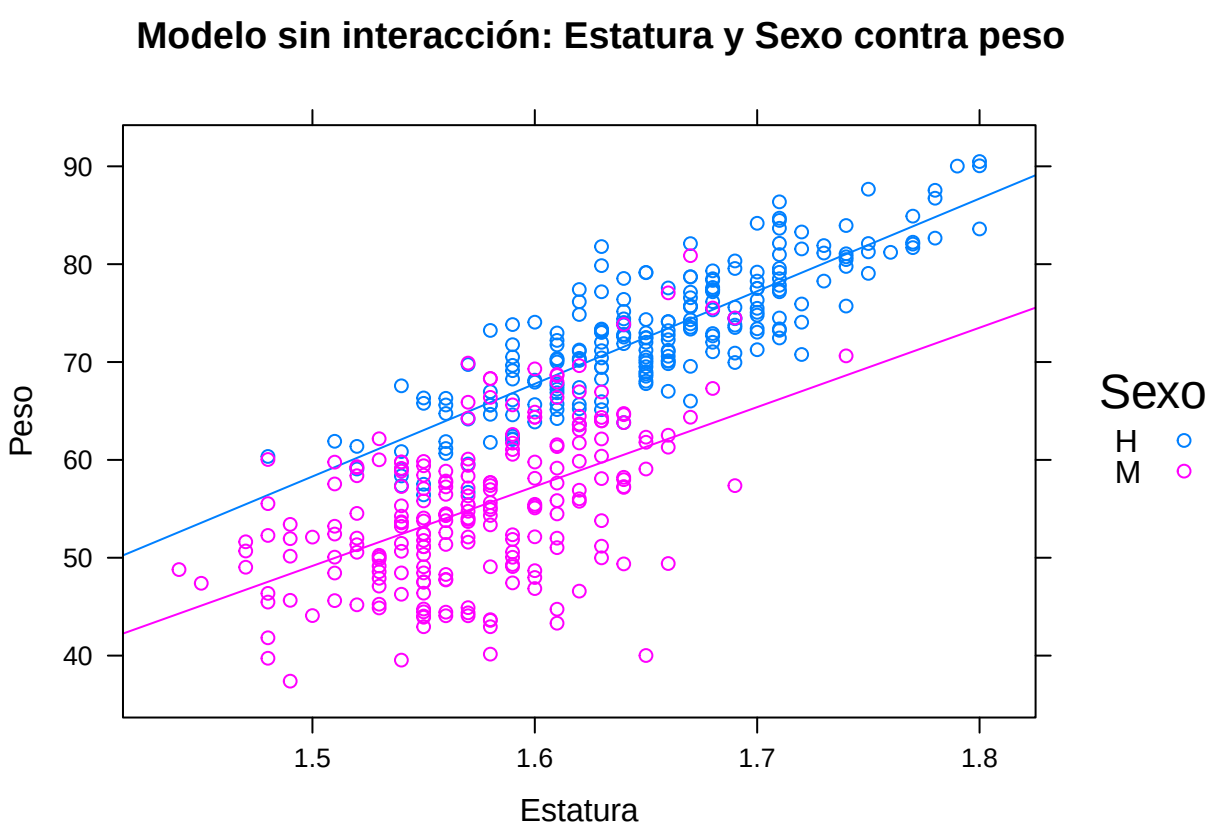
Conclusión mejor modelo

De esta manera, observamos que el mejor modelo es el de estatura sin interacción en sexo de todo el conjunto de datos, pues es significativo. Aunque tenemos el mayor coeficiente de determinación con un valor del 78.32% en el modelo con interacción, debido a que hay más variables el R^2 se infla, por lo que no necesariamente es el mejor.

Por lo anterior, nos quedamos con el modelo sin interacción, que este explica el 78.27% de la variación del peso de las personas.

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

```
library(lattice)
modelo <- lm(Peso ~ Estatura + Sexo, data = M)
xyplot(Peso ~ Estatura, data = M, groups = Sexo,
       auto.key = list(space = "right", title = "Sexo"),
       type = c("p", "r"),
       xlab = "Estatura", ylab = "Peso",
       main = "Modelo sin interacción: Estatura y Sexo contra peso")
```



Este modelo de regresión lineal multivariada logra el mejor ajuste del modelo, tomando en cuenta las variables de sexo y estatura sin interacción, logrando explicar el 78.27% de la variación del peso de las personas.

5 y 6. Interpreta en el contexto del problema cada uno de los análisis que hiciste.

El valor p representa en cada modelo si es significativo o no este. Es decir, que si hay una asociación real entre las variables y no debida al azar o por casualidad. Por otro lado, $SexoM$ representa si el sexo de la persona es mujer o no. El intercepto de esta variable representa el valor inicial que tomaría el modelo si es mujer (-10.56) y si no lo es (0). Indica que existe una relación negativa entre el sexo y el peso: Si es mujer, menor peso esperado. Finalmente la interacción Estatura:SexoM nos indica que el impacto de Estatura en el Peso es 13.51 unidades menor en las mujeres que en los hombres.

a. ¿Qué información proporciona $\hat{\beta}_0$ sobre la relación entre la estatura y el peso de hombres y mujeres?

β_0 es el intercepto y representa el peso cuando la estatura es cero. Es decir, el valor inicial que tiene el modelo y forma parte de la ecuación de regresión. El intercepto con mayor impacto es el de la interacción, pues influye en -83.68 unidades: cuando la estatura es 0, el peso es -83.68 (no tiene mucho sentido de manera literal, pero sí dentro del modelo de regresión lineal).

b. ¿Cómo interpretas $\hat{\beta}_i$ en la relación entre la estatura y el peso de hombres y mujeres?

β_1 representa el cambio en el peso por cada unidad de cambio en la estatura. Es decir, +1 m de estatura = + 35 kg en el peso. ∴ existe una relación positiva y directa entre la estatura y el peso: mayor estatura, mayor peso.

β_2 representa el cambio en el peso, dependiendo si es sexo. Es decir, si es mujer = 11.12 kg en el peso, si es hombre = +0 kg en el peso, en el modelo con interacción. Por otro lado, en el modelo sin interacción, si es mujer = -10.56kg en el peso, si es hombre = +0 kg en el peso.

Finalmente, β_3 representa el impacto de que tiene estatura, con respecto al sexo. Es decir, si es mujer = -13.51 kg en el peso, si es hombre = + 0 kg en el peso. Es decir, en el modelo con interacción nos quedaría que por cada unidad de estatura hay un cambio de 94.66-13.51 = 81.15 kg en el peso para mujeres, y un cambio de 94.66 kg en el peso para los hombres.

c. 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.

El modelo más apropiado es el de la relación entre peso y estatura sin interacción en el sexo, pues obtenemos un coeficiente de determinación del 78.27%, lo cual nos indica que el modelo explica el 78.27% de la variación del peso de las personas, y todas las variables que se toman en cuenta son estadísticamente significativas. En cambio, el modelo sin interacción no es apropriado, pues no hay una significancia en todas las variables, por lo que el R^2 solo es infla, debido a que hay una mayor cantidad de variables.

∴ No hay una interacción significativa entre la estatura y el sexo.

Análisis de los residuos

Modelo sin interacción entere estatura y sexo para predecir el peso

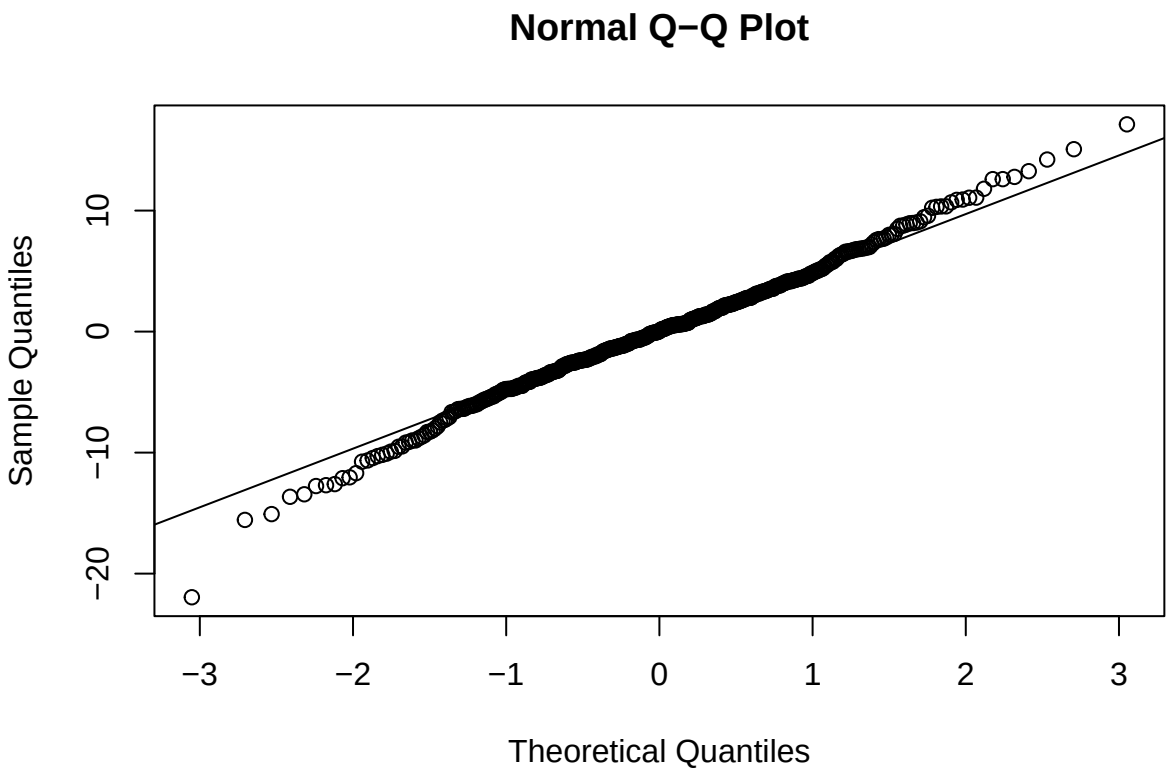
Normalidad

H_0 : Los residuos siguen una distribución normal. H_1 : Los residuos NO siguen una distribución normal.

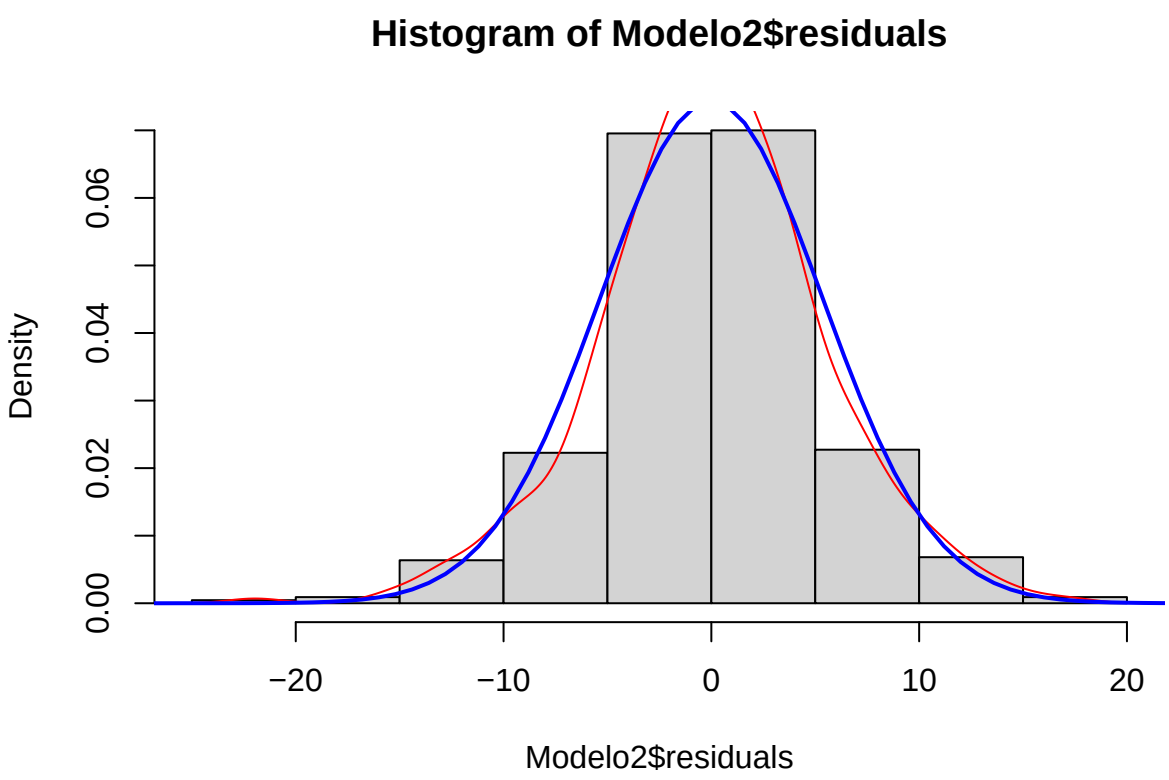
```
library(nortest)
ad.test(residuals(Modelo2))
```

```
##
##  Anderson-Darling normality test
##
## data:  residuals(Modelo2)
## A = 0.79651, p-value = 0.03879
```

```
qqnorm(Modelo2$residuals)
qqline(Modelo2$residuals)
```




```
hist(Modelo2$residuals,freq=FALSE)
lines(density(Modelo2$residuals),col="red")
curve(dnorm(x,mean=mean(Modelo2$residuals),sd=sd(Modelo2$residuals)), from=-
40, to=40, add=TRUE, col="blue",lwd=2)
```



Como podemos observar, el valor p > 0.03, por lo que H_0 no se rechaza y los residuos siguen una distribución normal.

Verificación de media cero

H_0: $\mu_e = 0$ H_1 : $\mu_e \neq 0$

```
t.test(Modelo2$residuals)
```

```
##
##  One Sample t-test
##
## data:  Modelo2$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
```

Como tenemos un valor p ≈ 1, H_0 no se rechaza, por lo que los residuos tienen media cero.

Homocedasticidad

H_0 : La varianza de los errores es constante (Hay homocedasticidad). H_1: La varianza de los errores NO es constannte (Hay heterocedasticidad).

```
library(lmtest)
```

```
## Loading required package: zoo
```

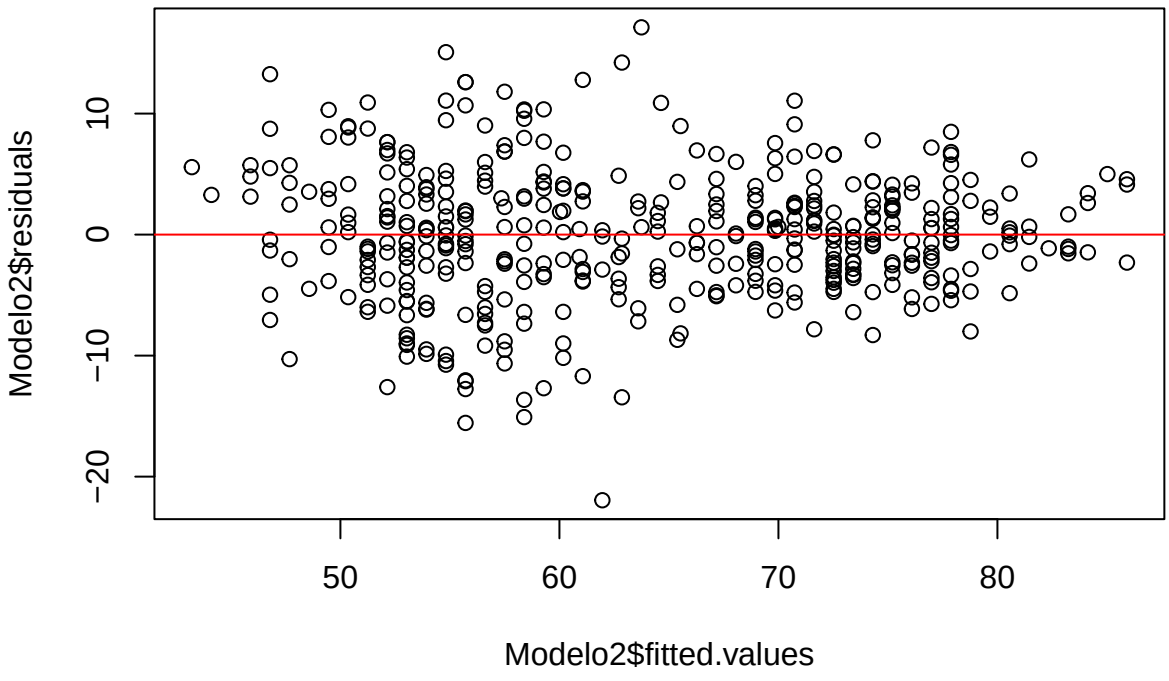
```
##
## Attaching package: 'zoo'
```

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

```
bptest(Modelo2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Modelo2
## BP = 48.202, df = 2, p-value = 3.413e-11
```

```
plot(Modelo2$fitted.values,Modelo2$residuals)
abline(h=0, col= 'red')
```



Como el valor p < 0.03, Se rechaza H_0, por lo que la varianza de los errores NO es constante (hay heterocedasticidad). Además, esto se puede observar en el gráfico, pues la varianza no fluctúa dentro de un rango, sino de manera caótica.

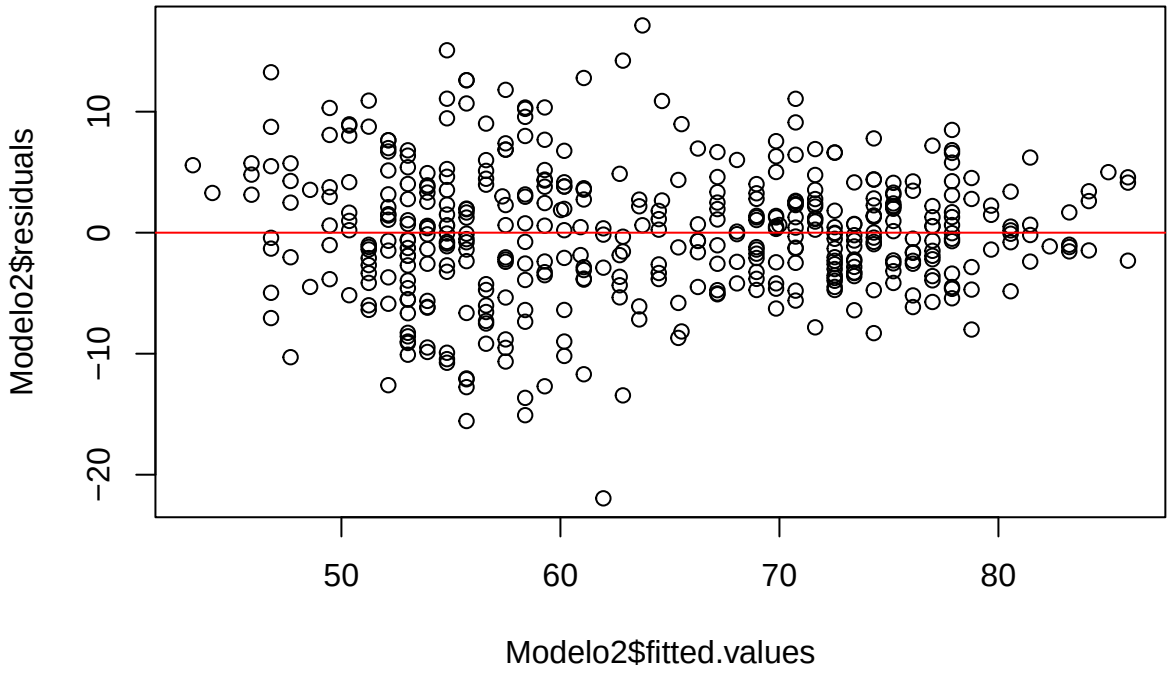
Independencia

H_0: La autocorrelación de los residuos es 0 (hay independencia). H_1: La autocorrelación de los residuos $\neq 0$ (no hay independencia).

```
dwtest(Modelo2)
```

```
##
##  Durbin-Watson test
##
## data:  Modelo2
## DW = 1.8663, p-value = 0.07325
## alternative hypothesis: true autocorrelation is greater than 0
```

```
plot(Modelo2$fitted.values,Modelo2$residuals)
abline(h=0, col= 'red')
```



Como el valor p > 0.03, no se rechaza H_0, por lo que sí hay independencia en los residuos. Además, no se observa un patrón en la gráfica, o una dependencia entre los residuos.

Conclusión

Debido a que el modelo no cumple con homocedasticidad, este modelo no es adecuado, por lo que compararemos con el modelo con interacción entre sexo y estatura para predecir el peso.

Modelo con interacción entre estatura y sexo para predecir el peso

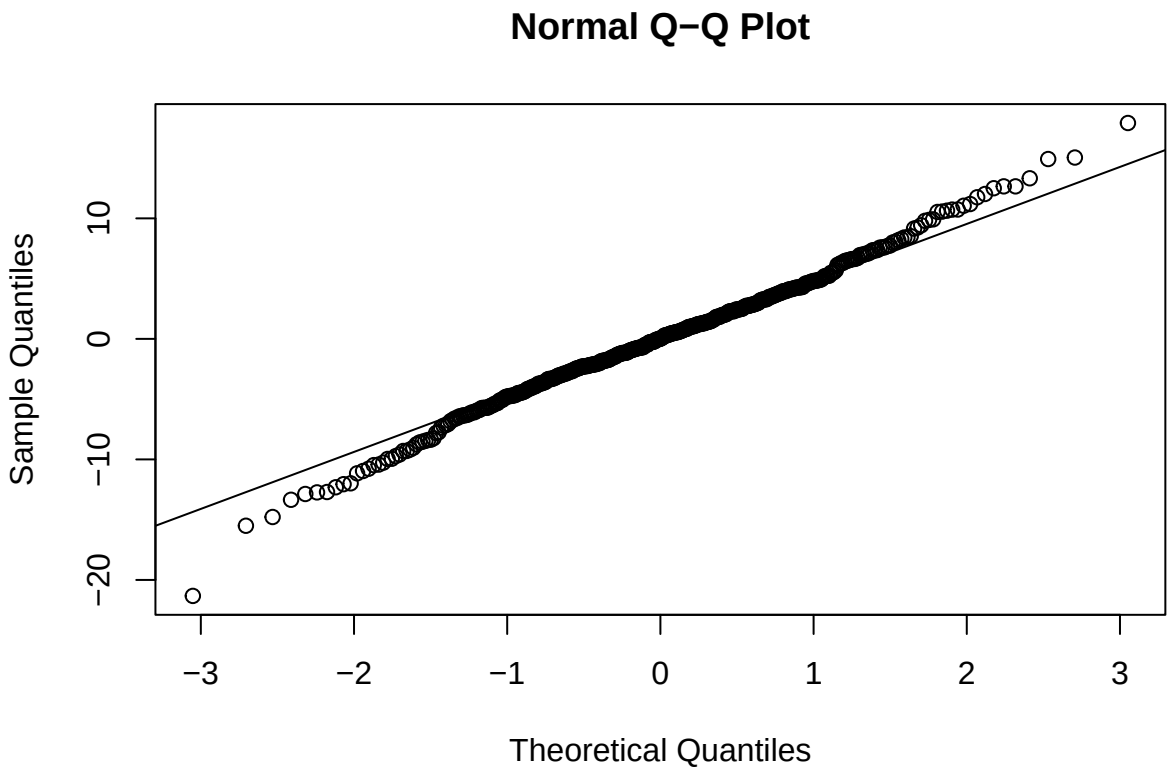
Normalidad

H_0 : Los residuos siguen una distribución normal. H_1 : Los residuos NO siguen una distribución normal.

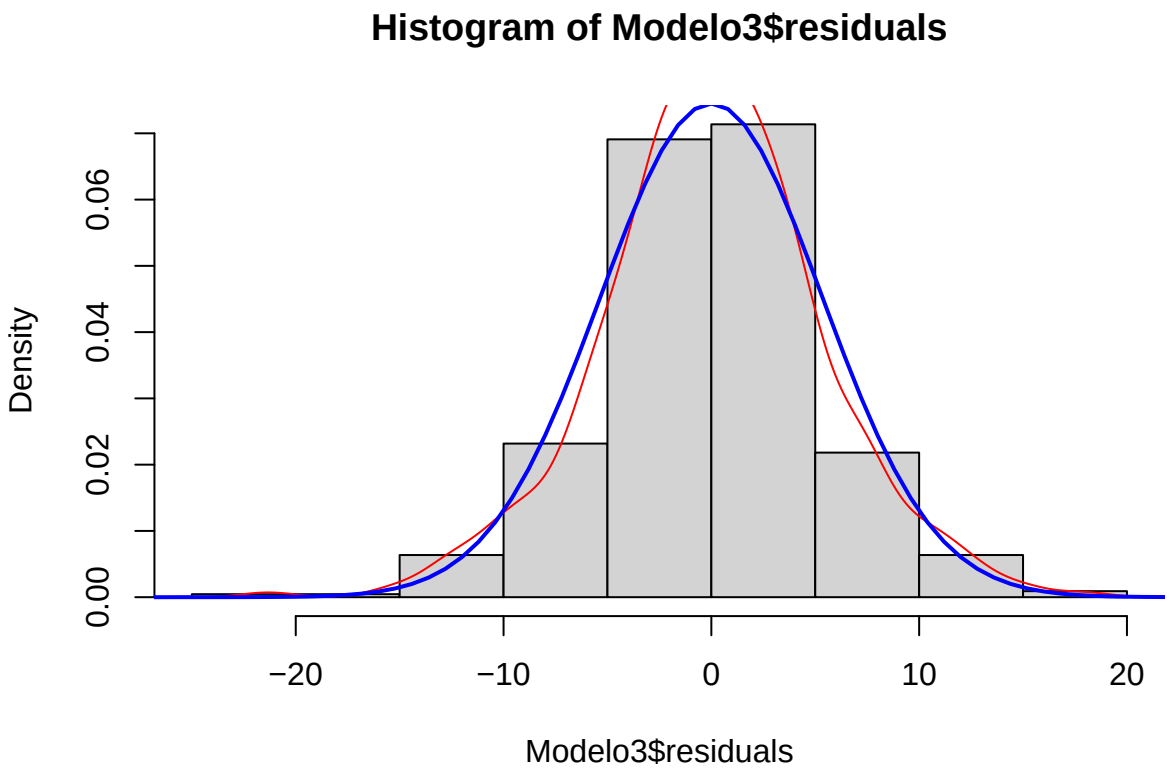
```
library(nortest)
ad.test(residuals(Modelo3))
```

```
##
##  Anderson-Darling normality test
##
## data:  residuals(Modelo3)
## A = 0.8138, p-value = 0.03516
```

```
qqnorm(Modelo3$residuals)
qqline(Modelo3$residuals)
```



```r
hist(Modelo3$residuals,freq=FALSE)
lines(density(Modelo3$residuals),col="red")
curve(dnorm(x,mean=mean(Modelo3$residuals),sd=sd(Modelo3$residuals)), from=-
40, to=40, add=TRUE, col="blue",lwd=2)
```



Como podemos observar, el valor p > 0.03, por lo que H_0 no se rechaza y los residuos siguen una distribución normal.

Verificación de media cero

H_0: $\mu_e = 0$ H_1 : $\mu_e \neq 0$

```r
t.test(Modelo3$residuals)
```

```
##
##  One Sample t-test
##
## data:  Modelo3$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
```

Como tenemos un valor p ≈ 1, H_0 no se rechaza, por lo que los residuos tienen media cero.

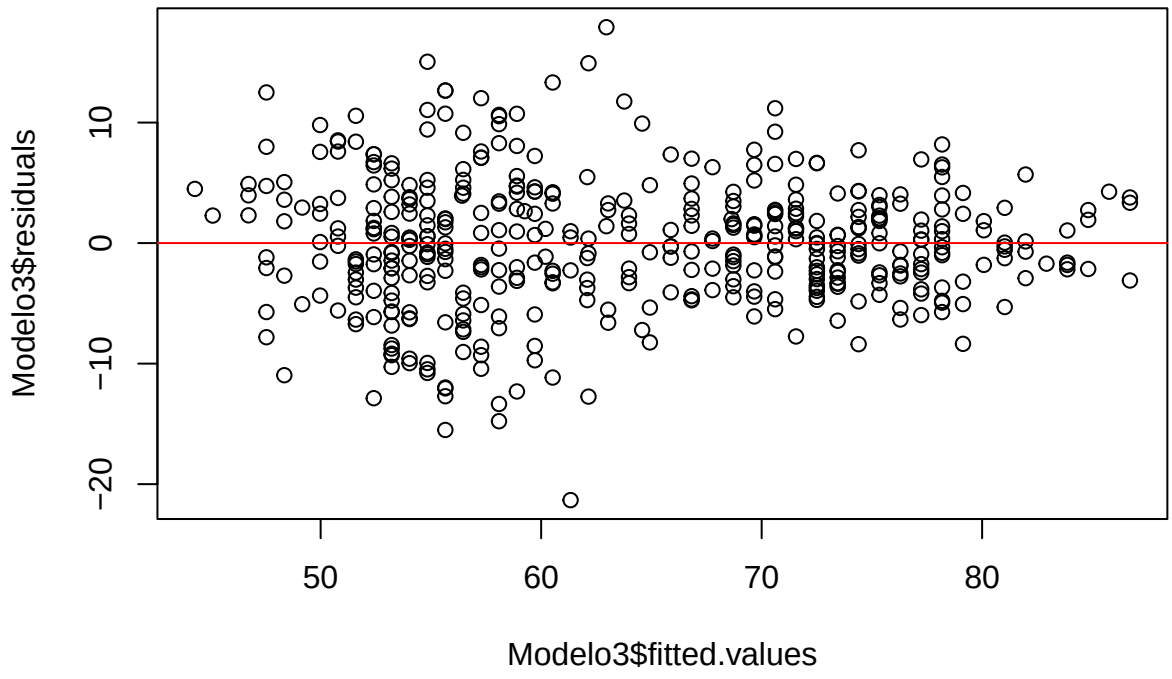
Homocedasticidad

H_0 : La varianza de los errores es constante (Hay homocedasticidad). H_1: La varianza de los errores NO es constannte (Hay heterocedasticidad).

```
library(lmtest)
bptest(Modelo3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Modelo3
## BP = 59.211, df = 3, p-value = 8.667e-13
```

```
plot(Modelo3$fitted.values,Modelo3$residuals)
abline(h=0, col= 'red')
```



Como el valor p < 0.03, Se rechaza H_0, por lo que la varianza de los errores NO es constante (hay heterocedasticidad). Además, esto se puede observar en el gráfico, pues se observa una variacion de los residuos muy desordenada.

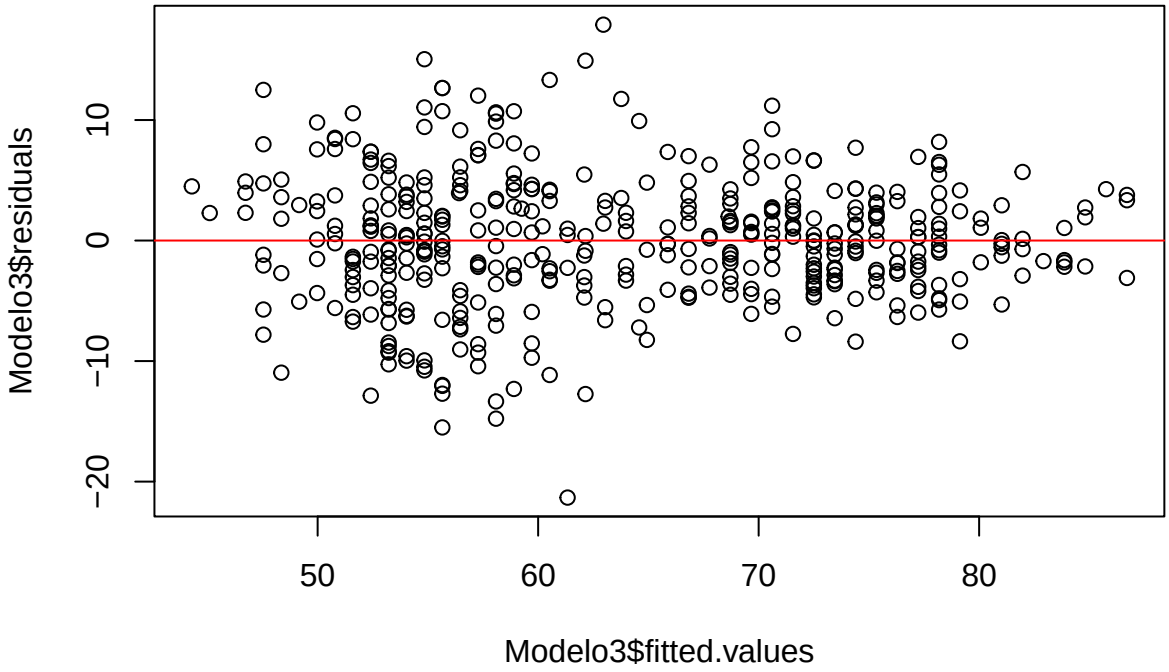
Independencia

H_0: La autocorrelación de los residuos es 0 (hay independencia). H_1: La autocorrelación de los residuos $\neq 0$ (no hay independencia).

```
dwtest(Modelo3)
```

```
##
##  Durbin-Watson test
##
## data:  Modelo3
## DW = 1.8646, p-value = 0.07113
## alternative hypothesis: true autocorrelation is greater than 0
```

```
plot(Modelo3$fitted.values,Modelo3$residuals)
abline(h=0, col= 'red')
```



Como el valor p > 0.03, no se rechaza H_0, por lo que sí hay independencia en los residuos. Además, no se observa un patrón en la gráfica, o una dependencia entre los residuos.

Conclusión

Debido a que el modelo no cumple con homocedasticidad, este modelo no es adecuado, por lo que compararemos con el modelo 1 que realizamos con estatura para predecir el peso, donde separamos los conjuntos de hombres y mujeres.

Modelo de regresión lineal de estatura para predecir el peso, por medio de la separación de conjuntos por sexo

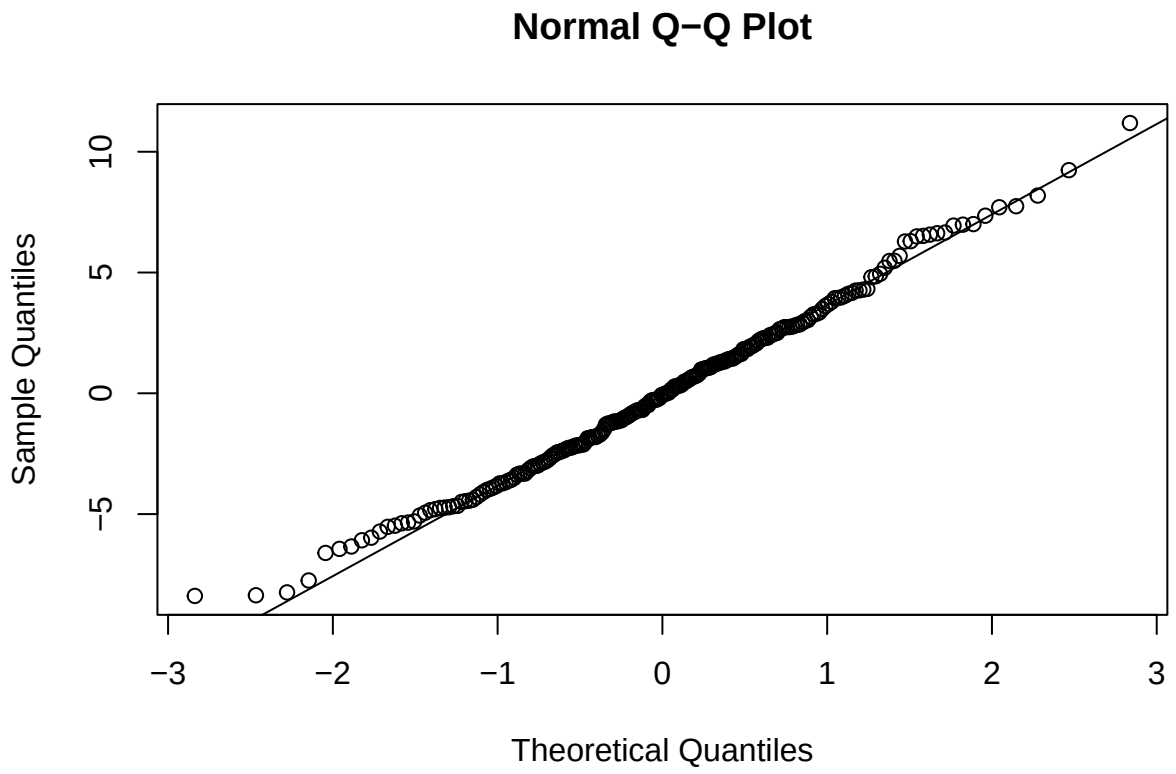
Normalidad

H_0 : Los residuos siguen una distribución normal. H_1 : Los residuos NO siguen una distribución normal.

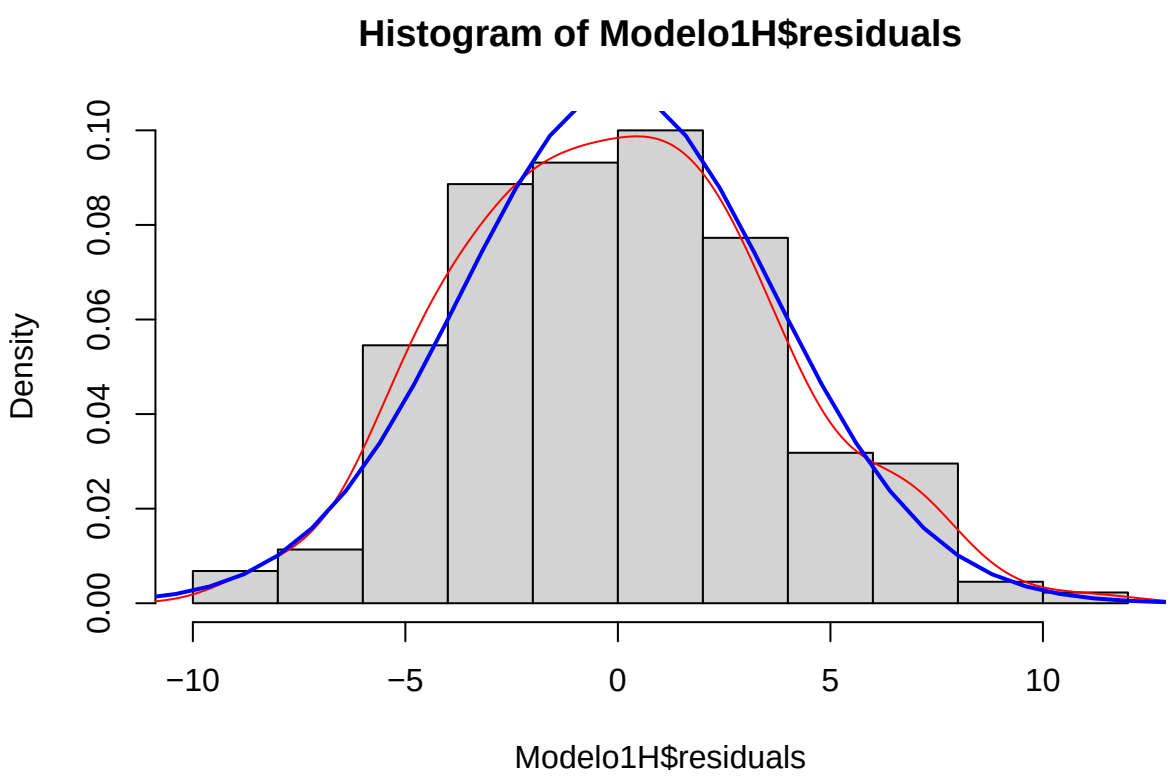
```
library(nortest)
ad.test(residuals(Modelo1H))
```

```
##
##  Anderson-Darling normality test
##
## data:  residuals(Modelo1H)
## A = 0.3009, p-value = 0.5771
```

```
qqnorm(Modelo1H$residuals)
qqline(Modelo1H$residuals)
```



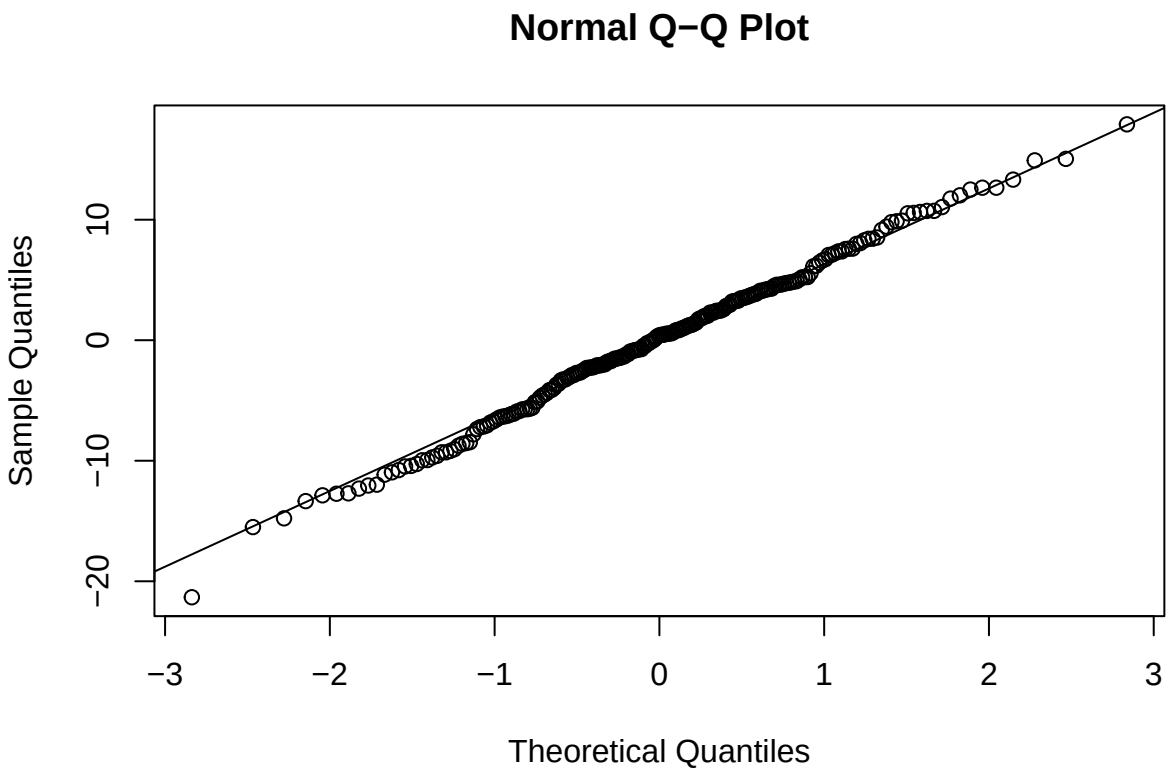

```
hist(Modelo1H$residuals,freq=FALSE)
lines(density(Modelo1H$residuals),col="red")
curve(dnorm(x,mean=mean(Modelo1H$residuals),sd=sd(Modelo1H$residuals)), from=-
40, to=40, add=TRUE, col="blue",lwd=2)
```



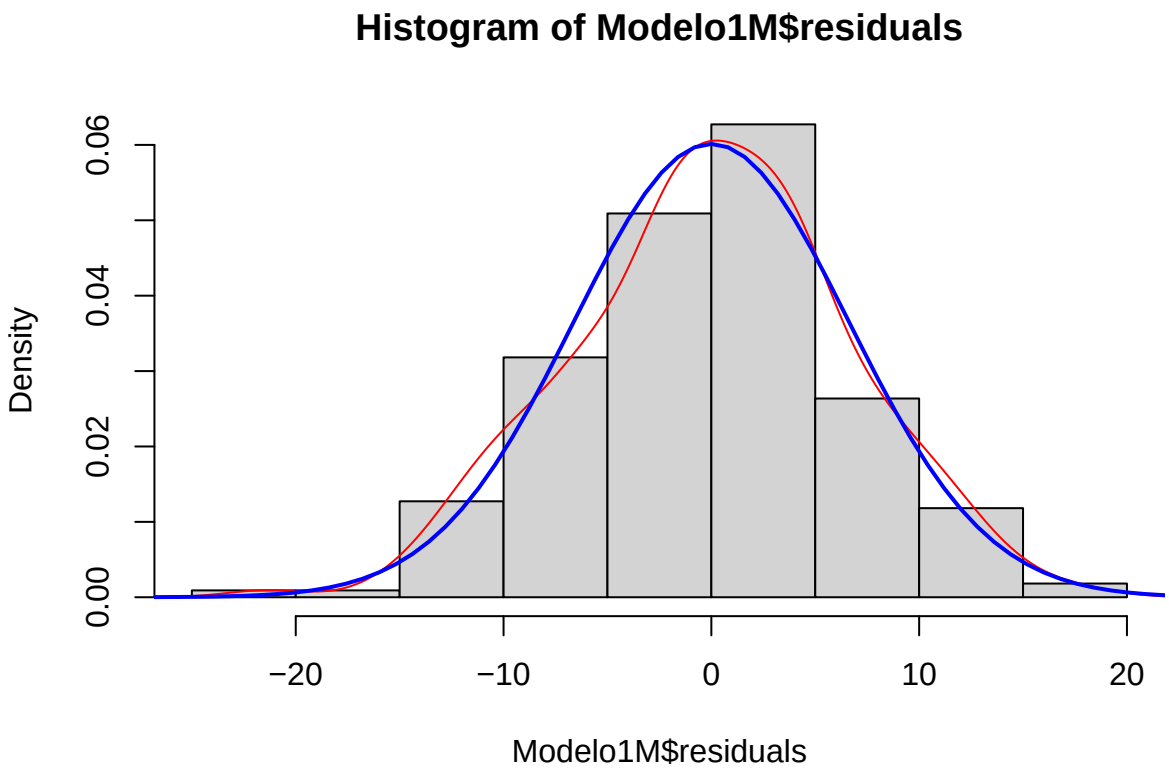

```
library(nortest)
ad.test(residuals(Modelo1M))
```

```
##
##  Anderson-Darling normality test
##
## data:  residuals(Modelo1M)
## A = 0.24899, p-value = 0.7451
```

```
qqnorm(Modelo1M$residuals)
qqline(Modelo1M$residuals)
```




```
hist(Modelo1M$residuals,freq=FALSE)
lines(density(Modelo1M$residuals),col="red")
curve(dnorm(x,mean=mean(Modelo1M$residuals),sd=sd(Modelo1M$residuals)), from=-
40, to=40, add=TRUE, col="blue",lwd=2)
```



Como podemos observar, los valores p $>$ 0.03, por lo que H_0 no se rechaza y los residuos siguen una distribución normal para hombres y mujeres.

Verificación de media cero

H_0: $\mu_e = 0$ H_1 : $\mu_e \neq 0$

```
t.test(Modelo1H$residuals)
```

```
##
##  One Sample t-test
##
## data:  Modelo1H$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
```

```
t.test(Modelo1M$residuals)
```

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

```
##
## data:  Modelo1M$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
```

Como tenemos un valor p ≈ 1, H_0 no se rechaza, por lo que los residuos tienen media cero para hombres y mujeres.

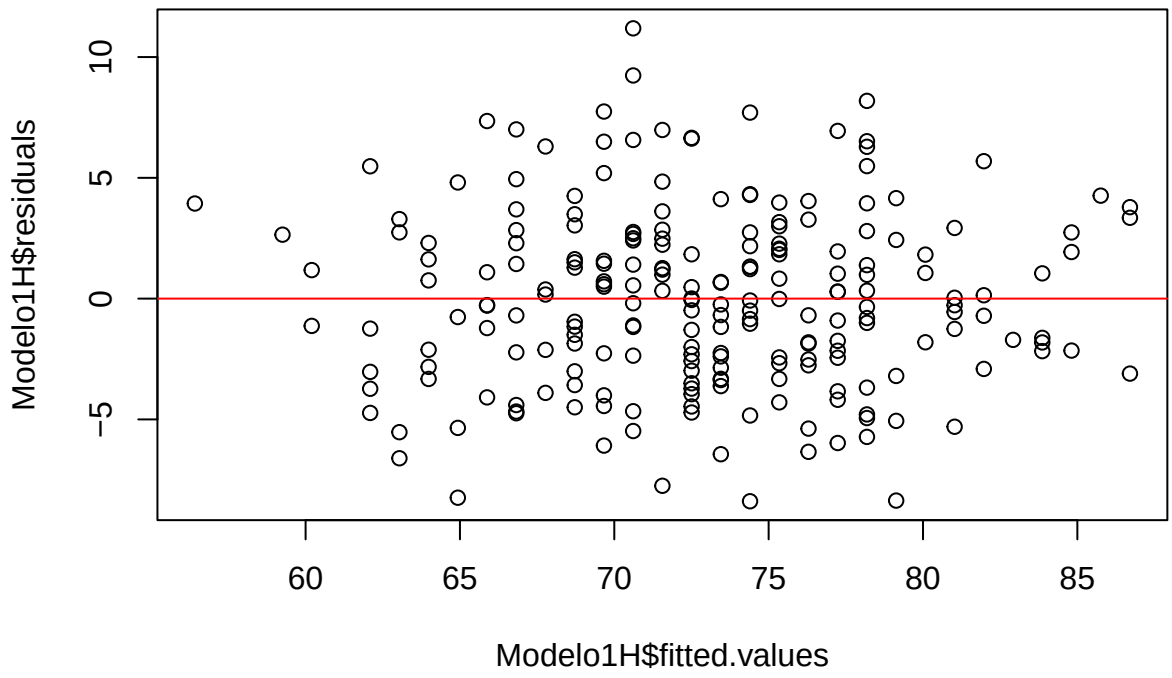
Homocedasticidad

H_0 : La varianza de los errores es constante (Hay homocedasticidad). H_1: La varianza de los errores NO es constannte (Hay heterocedasticidad).

```
library(lmtest)
bptest(Modelo1H)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Modelo1H
## BP = 0.93324, df = 1, p-value = 0.334
```

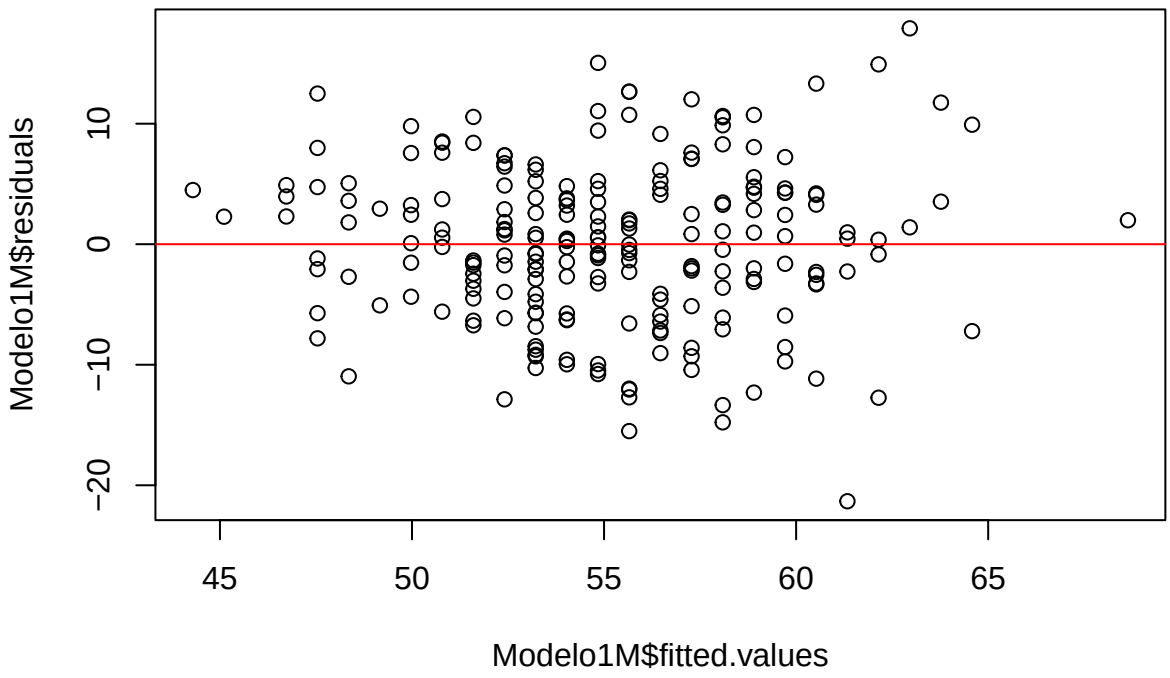
```
plot(Modelo1H$fitted.values,Modelo1H$residuals)
abline(h=0, col= 'red')
```




```
library(lmtest)
bptest(Modelo1M)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Modelo1M
## BP = 8.4976, df = 1, p-value = 0.003556
```

```
plot(Modelo1M$fitted.values,Modelo1M$residuals)
abline(h=0, col= 'red')
```



Como se puede observar, el valor p > 0.03 para el conjunto de hombres, por lo que sí hay homocedasticidad en el residuo de los hombres. Por otro lado, el valor p < 0.03 para el conjunto de mujeres, por lo que no hay homocedasticidad en el residuo de los mujeres. Debido a que la homocedasticidad no se cumple para ambos casos, no podemos validar este supuesto, por lo que se concluye que este modelo no tiene homocedasticidad.

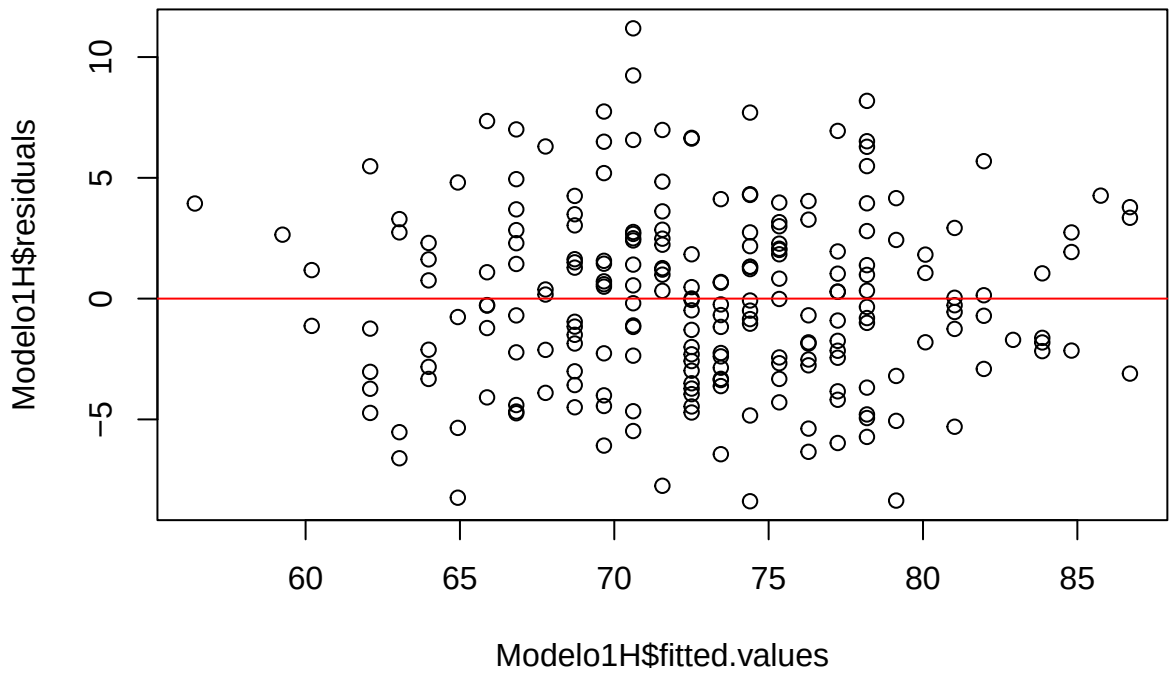
Independencia

H_0: La autocorrelación de los residuos es 0 (hay independencia). H_1: La autocorrelación de los residuos $\neq 0$ (no hay independencia).

```
dwtest(Modelo1H)
```

```
##
##  Durbin-Watson test
##
## data:  Modelo1H
## DW = 2.0556, p-value = 0.6599
## alternative hypothesis: true autocorrelation is greater than 0
```

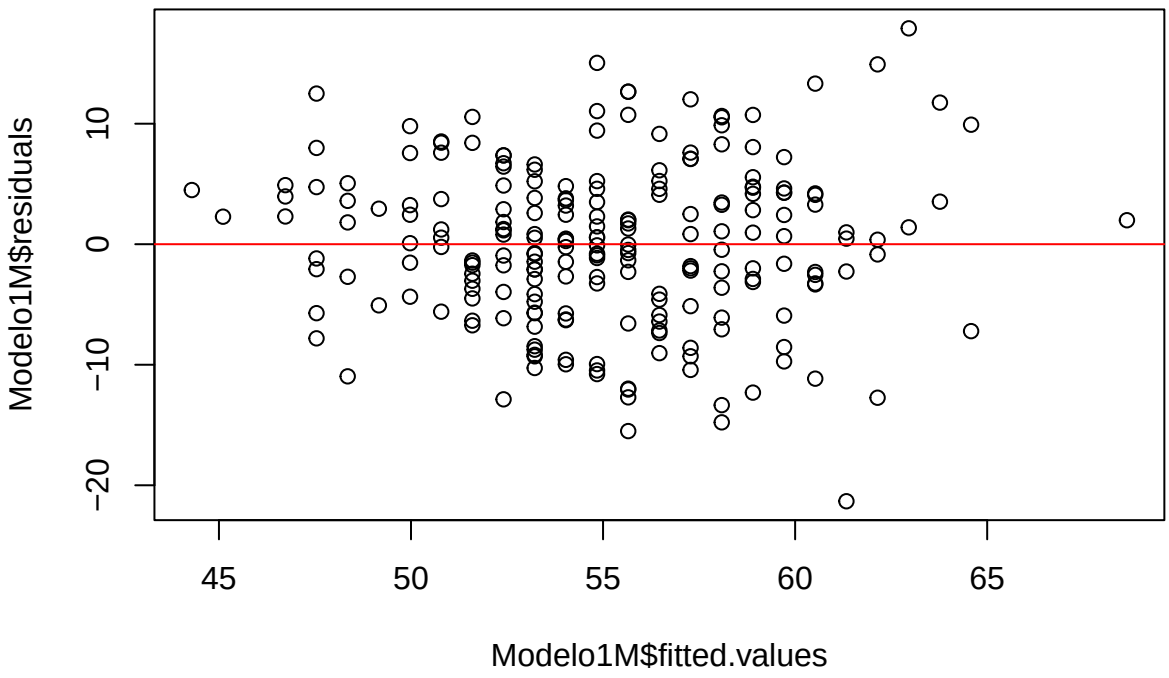
```
plot(Modelo1H$fitted.values,Modelo1H$residuals)
abline(h=0, col= 'red')
```




```
dwtest(Modelo1M)
```

```
##
##  Durbin-Watson test
##
## data:  Modelo1M
## DW = 1.8062, p-value = 0.07532
## alternative hypothesis: true autocorrelation is greater than 0
```

```
plot(Modelo1M$fitted.values,Modelo1M$residuals)
abline(h=0, col= 'red')
```



Como el valor p > 0.03, no se rechaza H_0, por lo que sí hay independencia en los residuos. Además, no se observa un patrón en la gráfica, o una dependencia entre los residuos.

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

La normalidad nos indica que la distribución de los pesos conforme a la estatura en los modelos de regresión lineal es normal. Es decir, la media, mediana y moda son cercanas, y su curtosis y sesgo son adecuados.

La independencia nos indica que el valor del peso con respecto a la estatura no depende del valor anterior. Es decir, no hay una secuencia temporal como una serie de tiempo.

La variación de media cero nos indica que los errores entre el valor estimado y observado son estadísticamente indistintos entre sí.

Finalmente, la homocedasticidad nos indica que la variación de los residuos es constante. Como no se encuentra en los modelos, quiere decir que el error entre peso observado y el peso predicho no puede ser determinado con un rango, pues este no es constante.

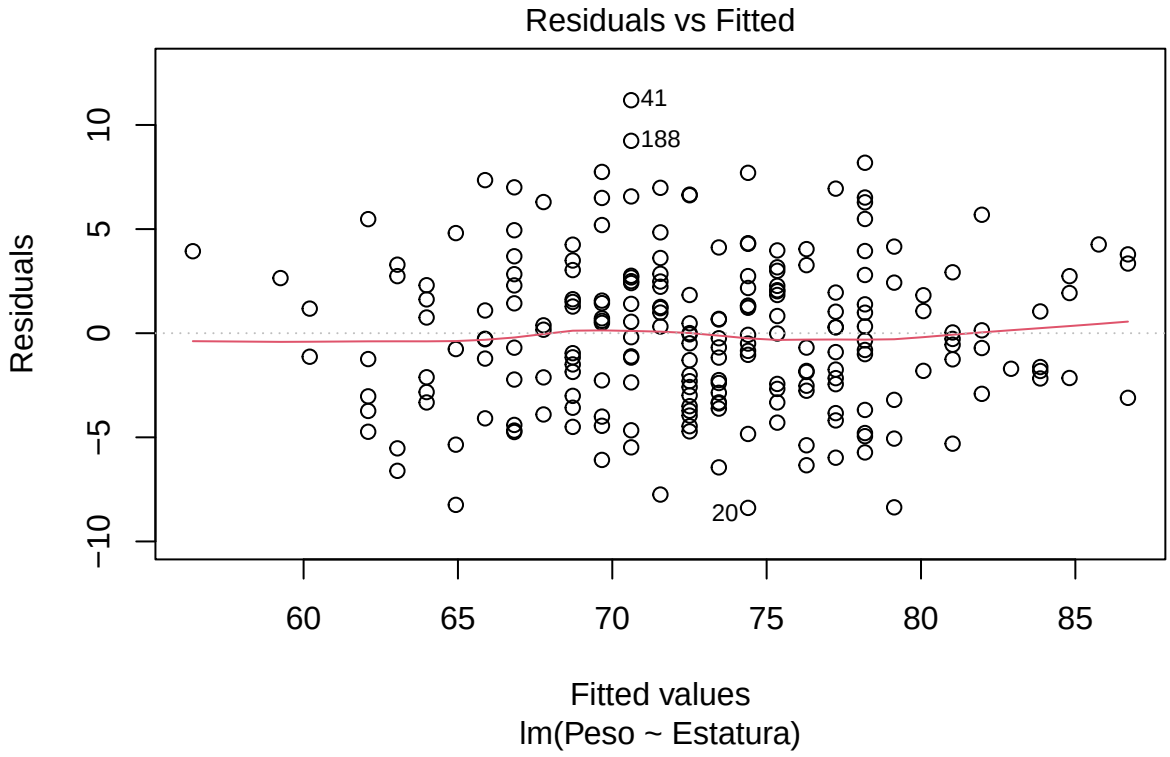
Conclusión final

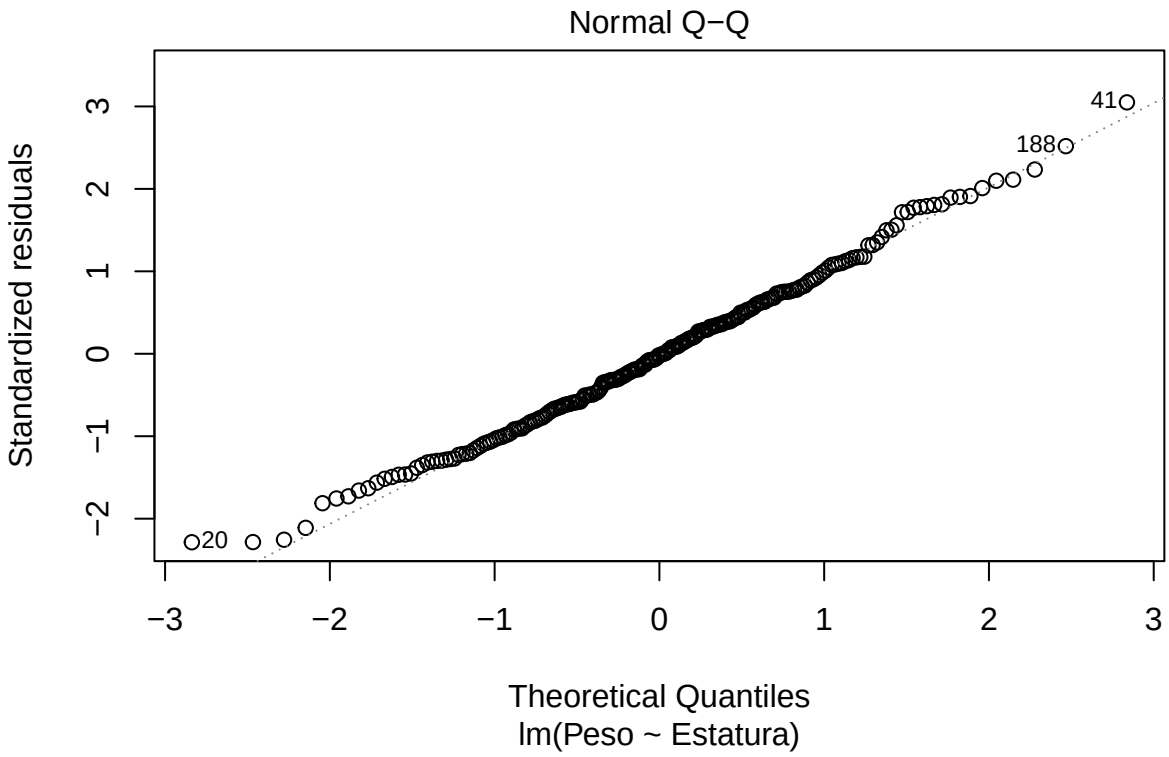
Los modelos de estatura y sexo con interacción y sin interacción para predecir el peso cumplen todos los supuestos, excepto del de homocedasticidad, por lo que no se pueden considerar adecuados. Además, el modelo en el que separamos el sexo en dos conjuntos diferentes y solo usamos la estatura para predecir validó todos los supuestos para el conjunto de hombres, pero no pudo validar la homocedasticidad para

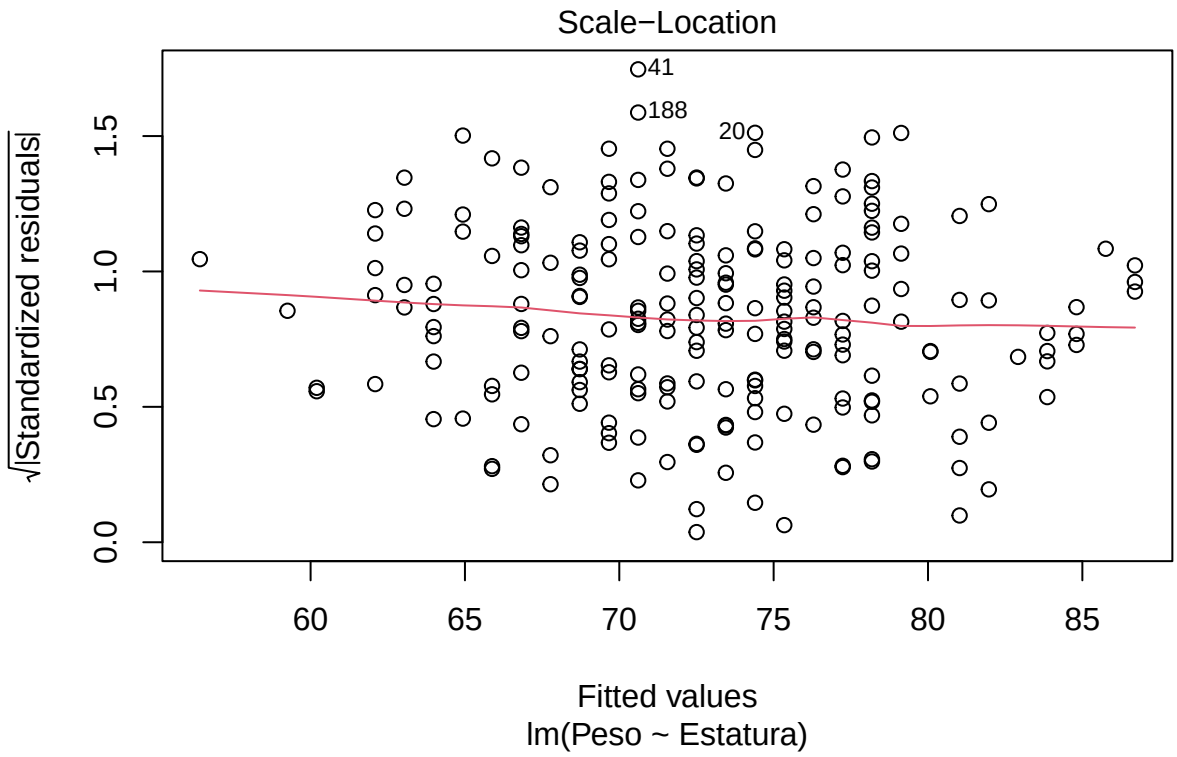
el conjunto de mujeres. Sin embargo, este modelo es el que estuvo más cerca de cumplir el supuesto de homocedasticidad, pues lo validó para hombres y tuvo un valor p de 0.0036 para la varianza de los residuos para mujeres, por lo que se puede concluir que es el modelo más cercano a ser validado y el mejor (o menos peor).

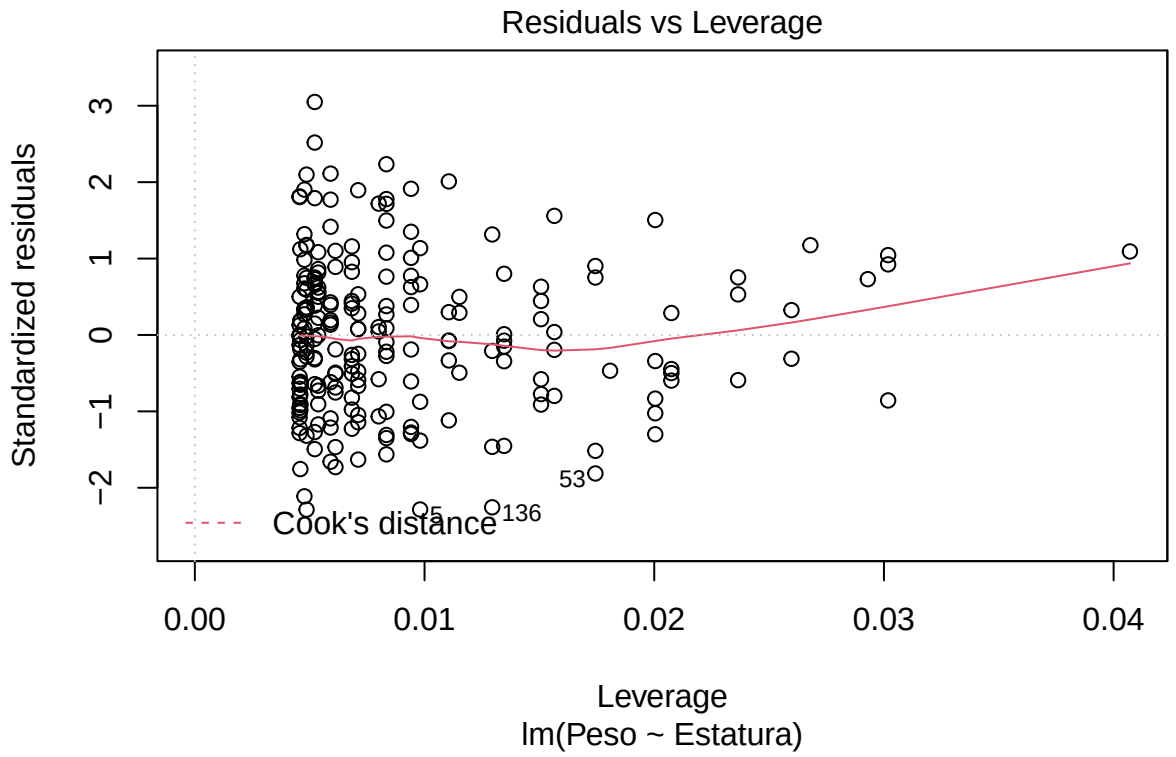
4. Utiliza el comando: plot(modelo). Observa las gráficas obtenidas y contesta:

```
plot(Modelo1H)
```

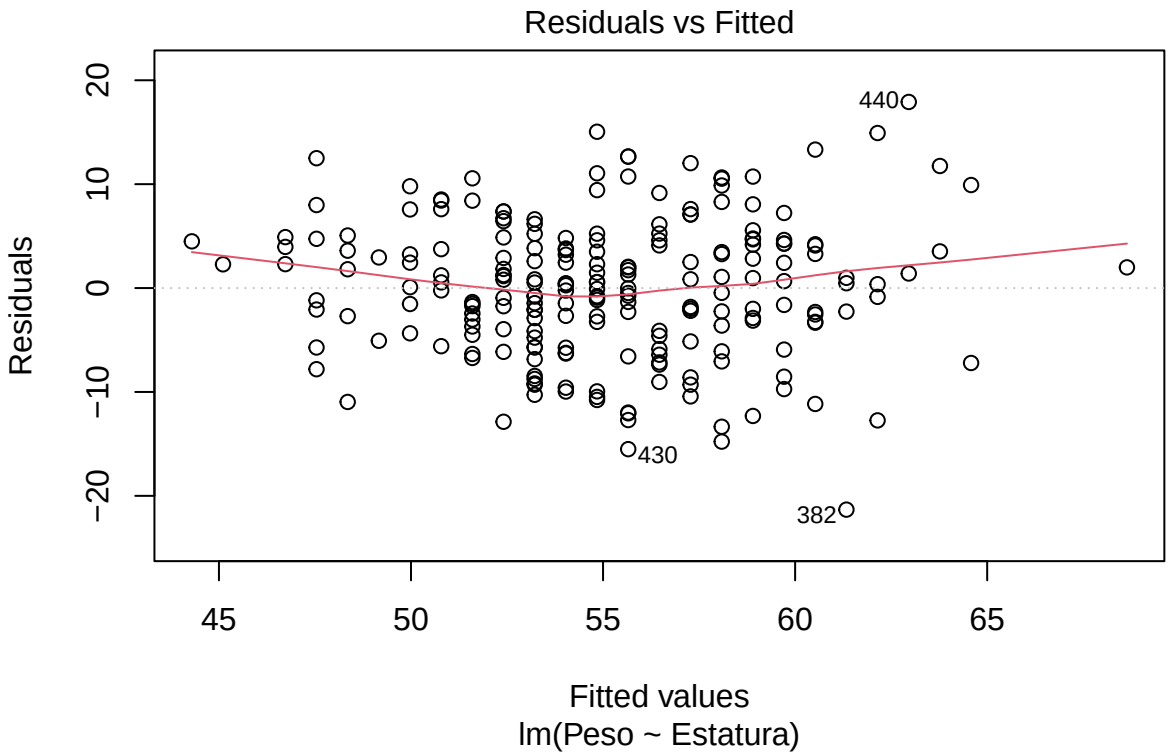


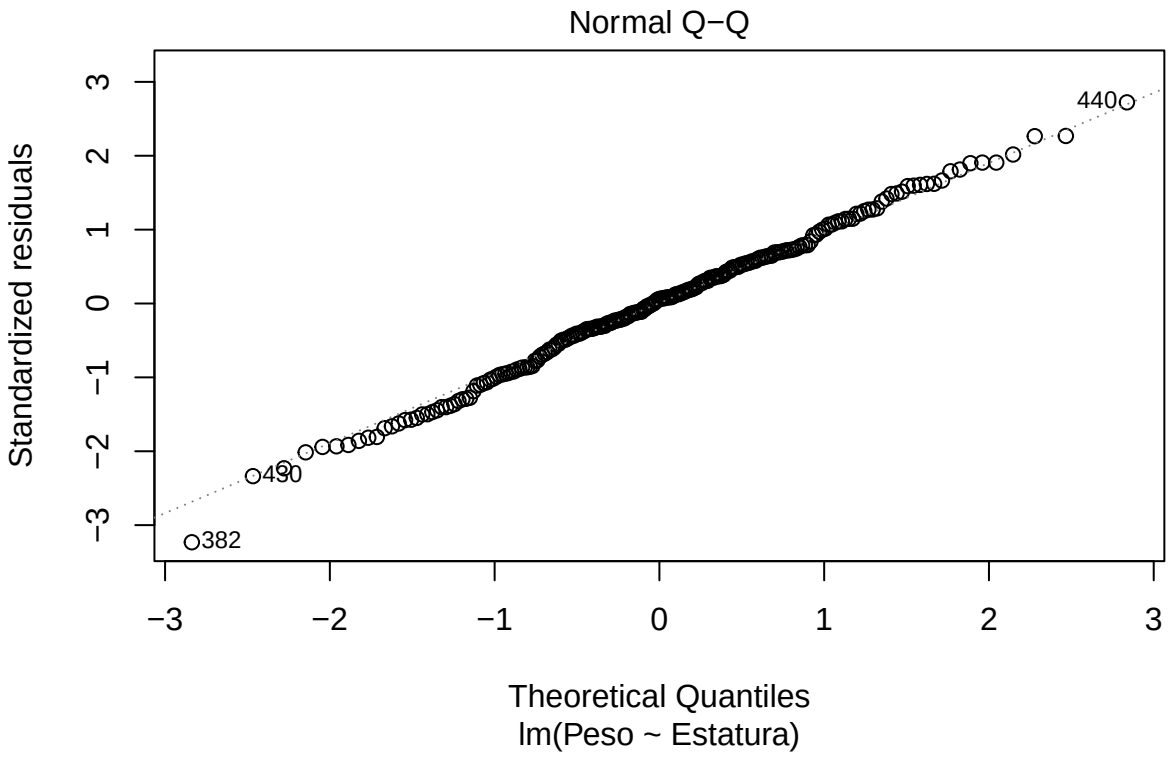

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Peso ~ Estatura)
41
188
20
−3
−2
−1
0
1
2
3

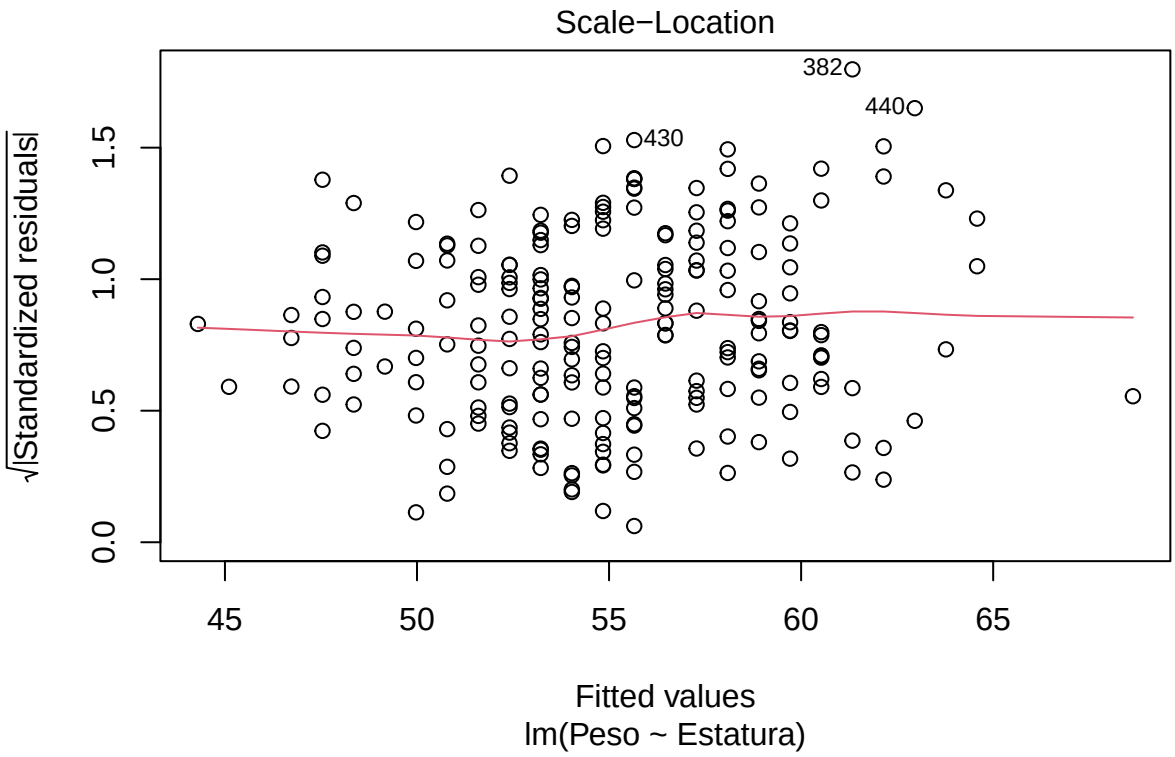

Scale−Location
41
188
20
√|Standardized residuals|
0.0
0.5
1.0
1.5
60
65
70
75
80
85
Fitted values
lm(Peso ~ Estatura)

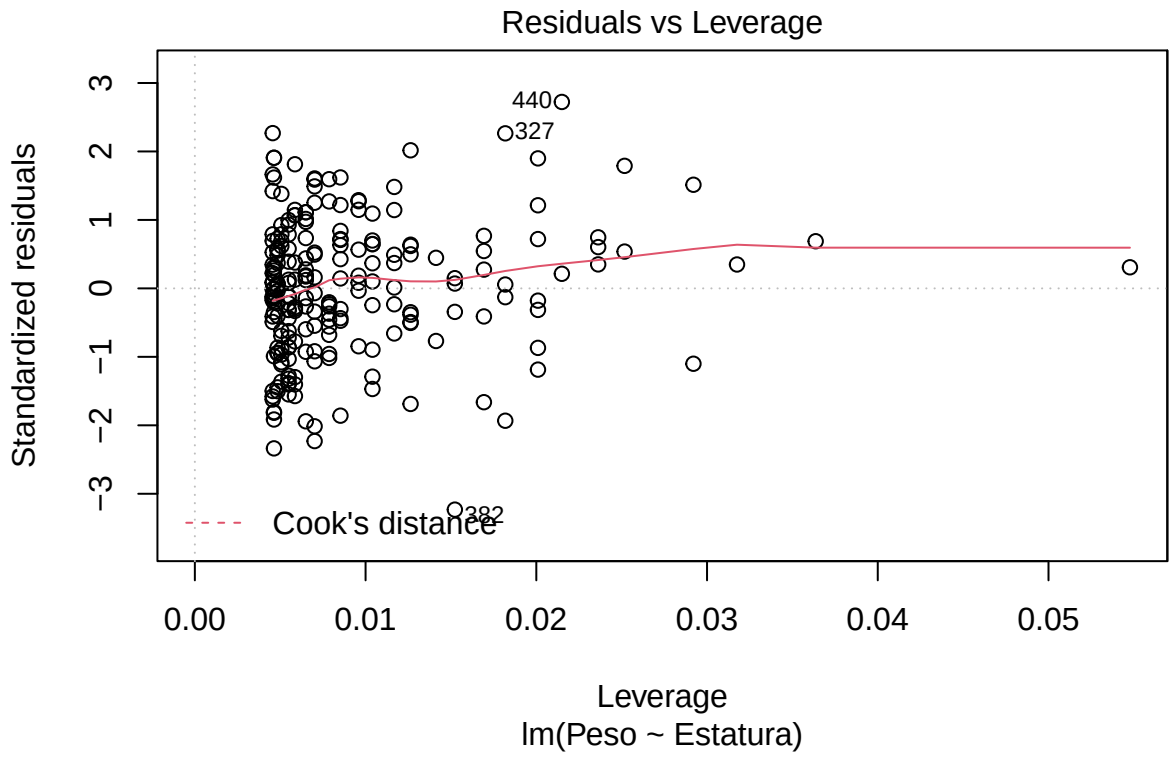



```
plot(Modelo1M)
```

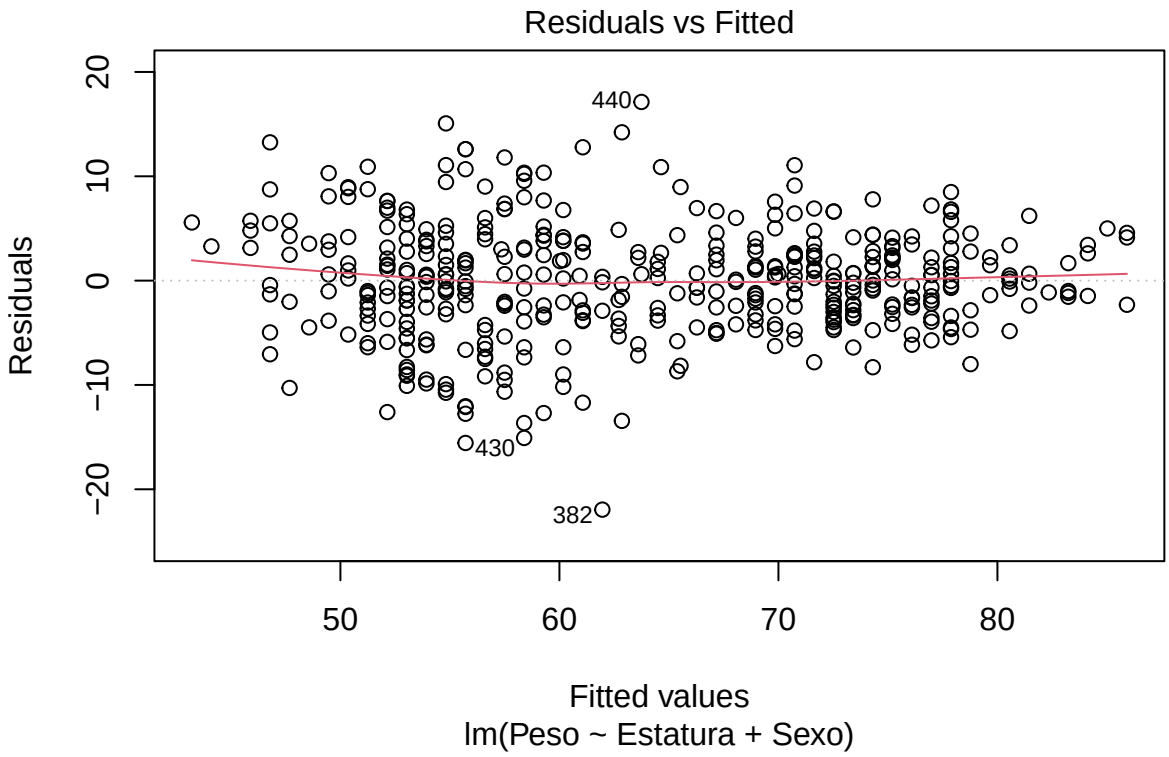

Residuals vs Fitted
Residuals
20
10
0
−10
−20
440
430
382
45
50
55
60
65
Fitted values
lm(Peso ~ Estatura)

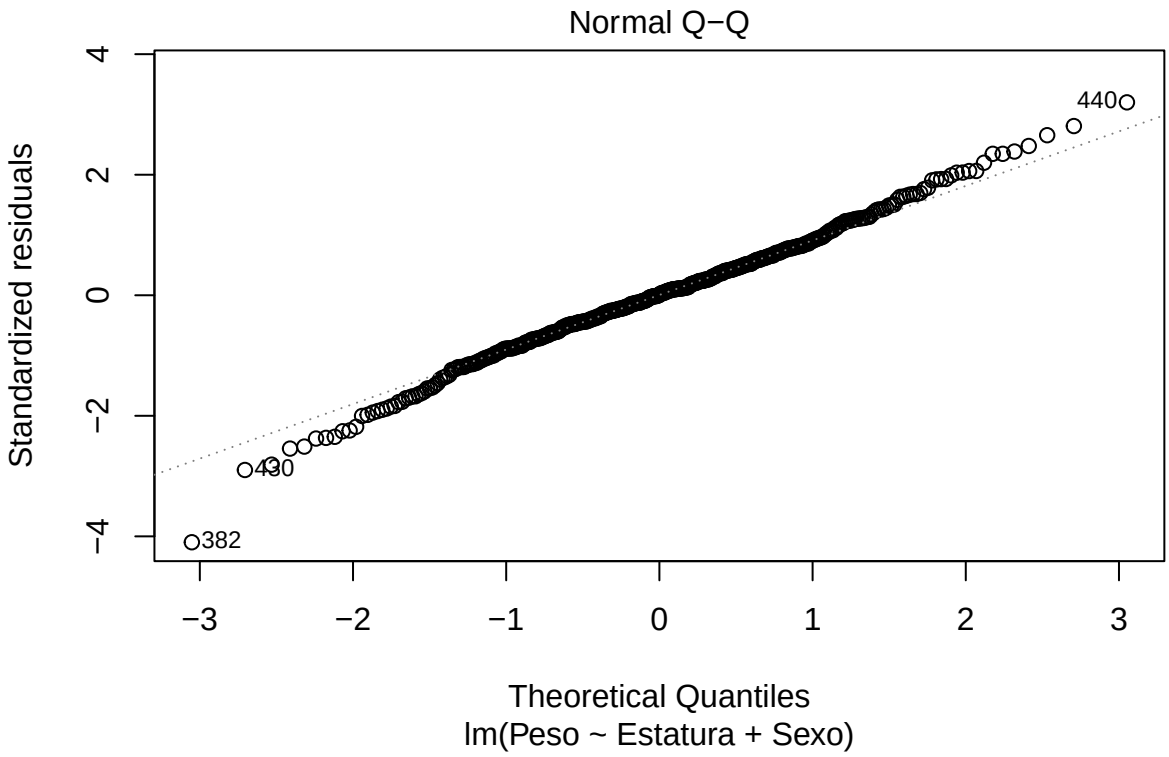

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Peso ~ Estatura)
440
490
382
−3
−2
−1
0
1
2
3

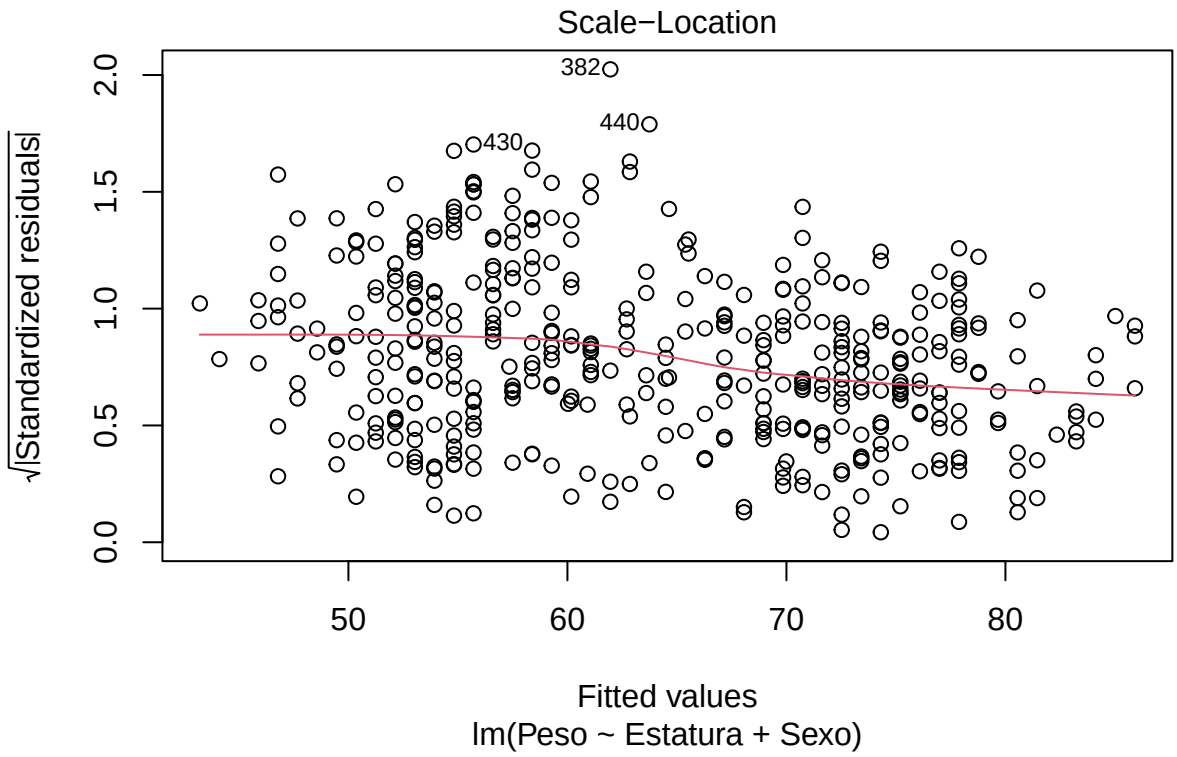

Scale−Location
382
440
430
√|Standardized residuals|
0.0
0.5
1.0
1.5
45
50
55
60
65
Fitted values
lm(Peso ~ Estatura)

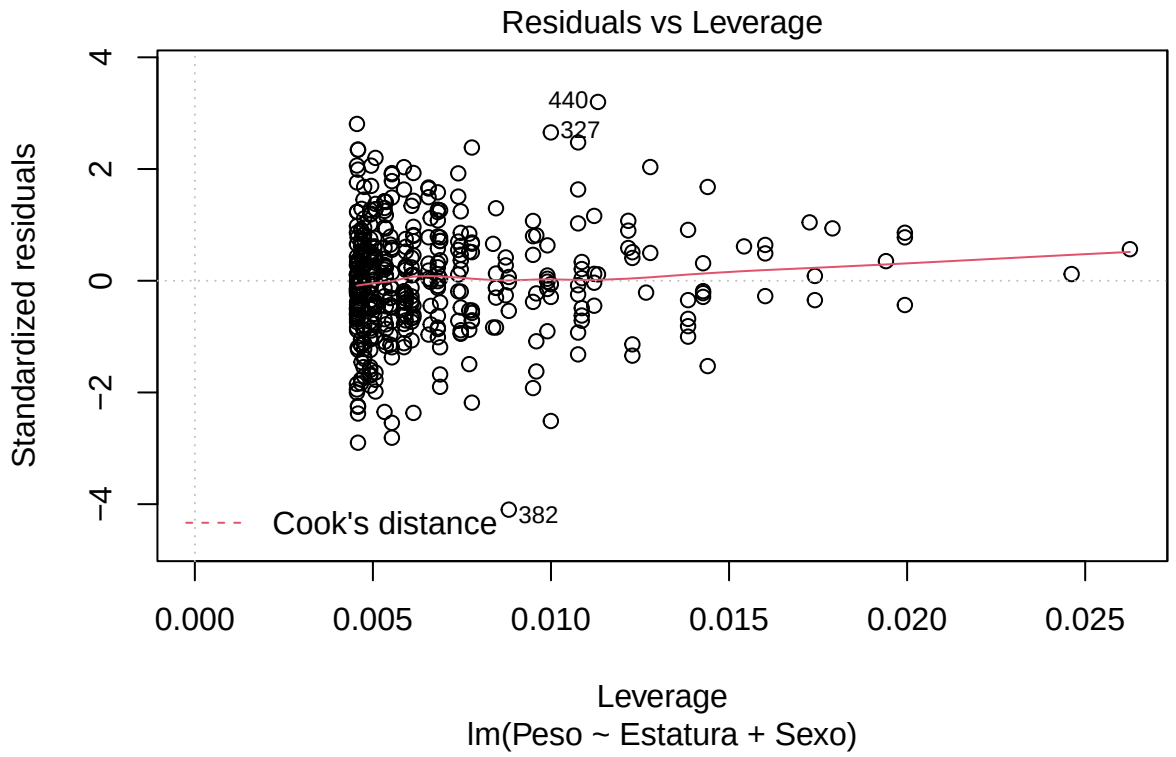



```
plot(Modelo2)
```

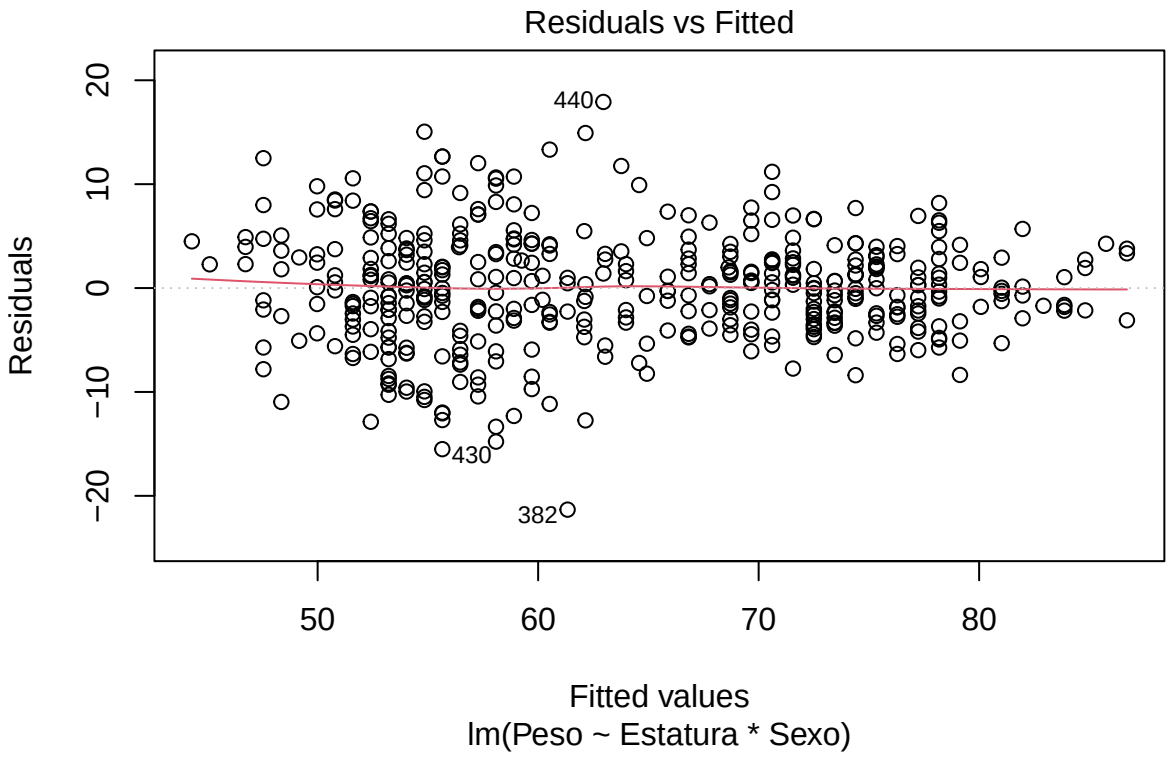

Residuals vs Fitted
Residuals
20
10
0
−10
−20
440
430
382
50
60
70
80
Fitted values
lm(Peso ~ Estatura + Sexo)

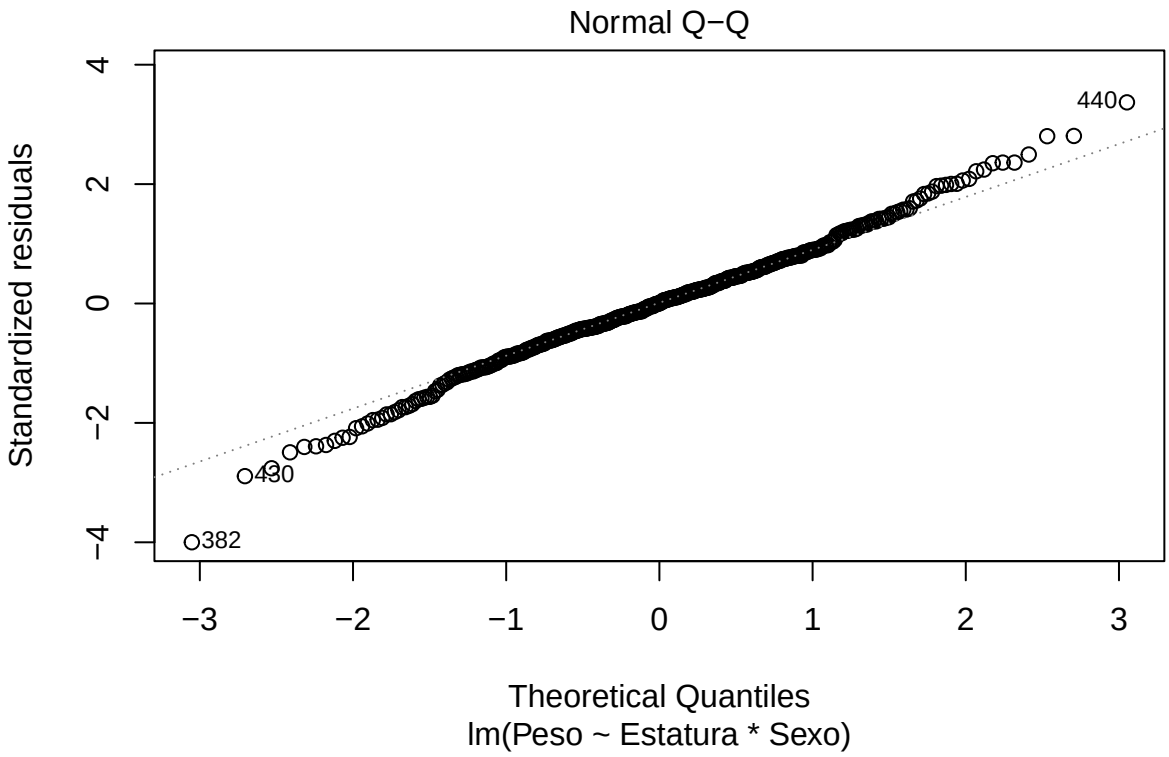

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Peso ~ Estatura + Sexo)
440
430
382
4
2
0
−2
−4
−3
−2
−1
0
1
2
3

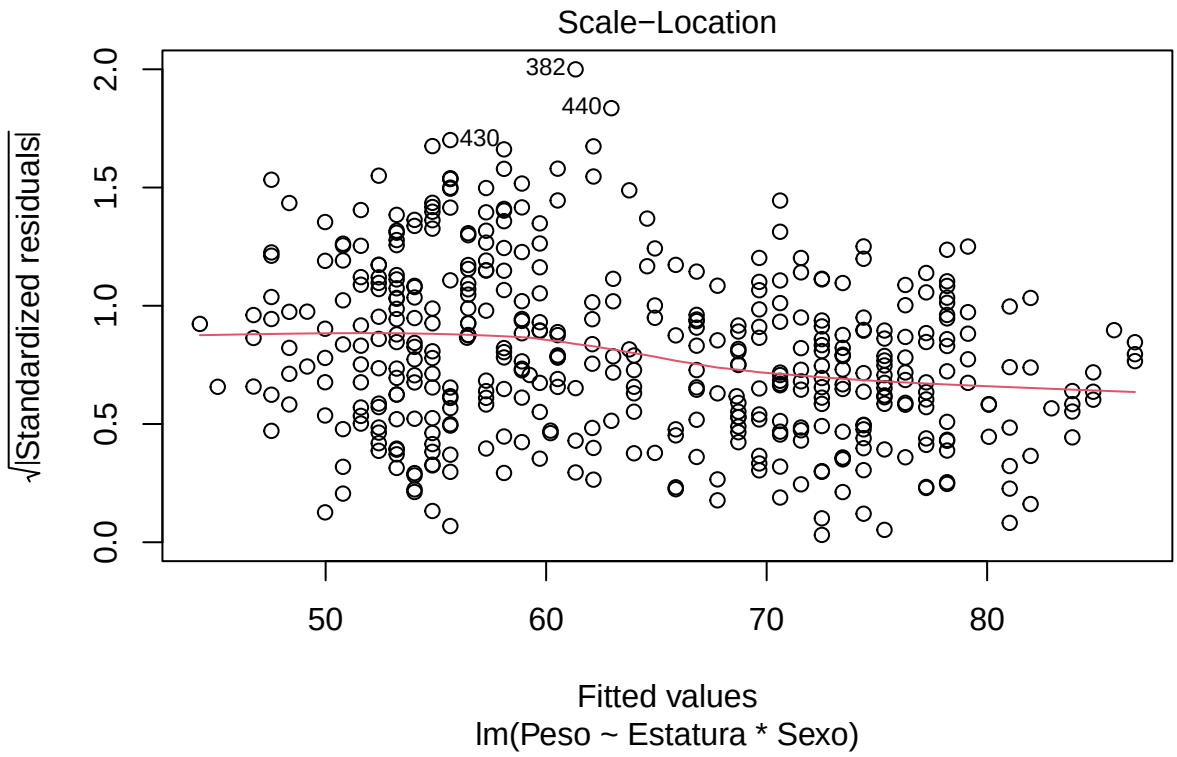

Scale−Location
382
440
430
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
50
60
70
80
Fitted values
lm(Peso ~ Estatura + Sexo)

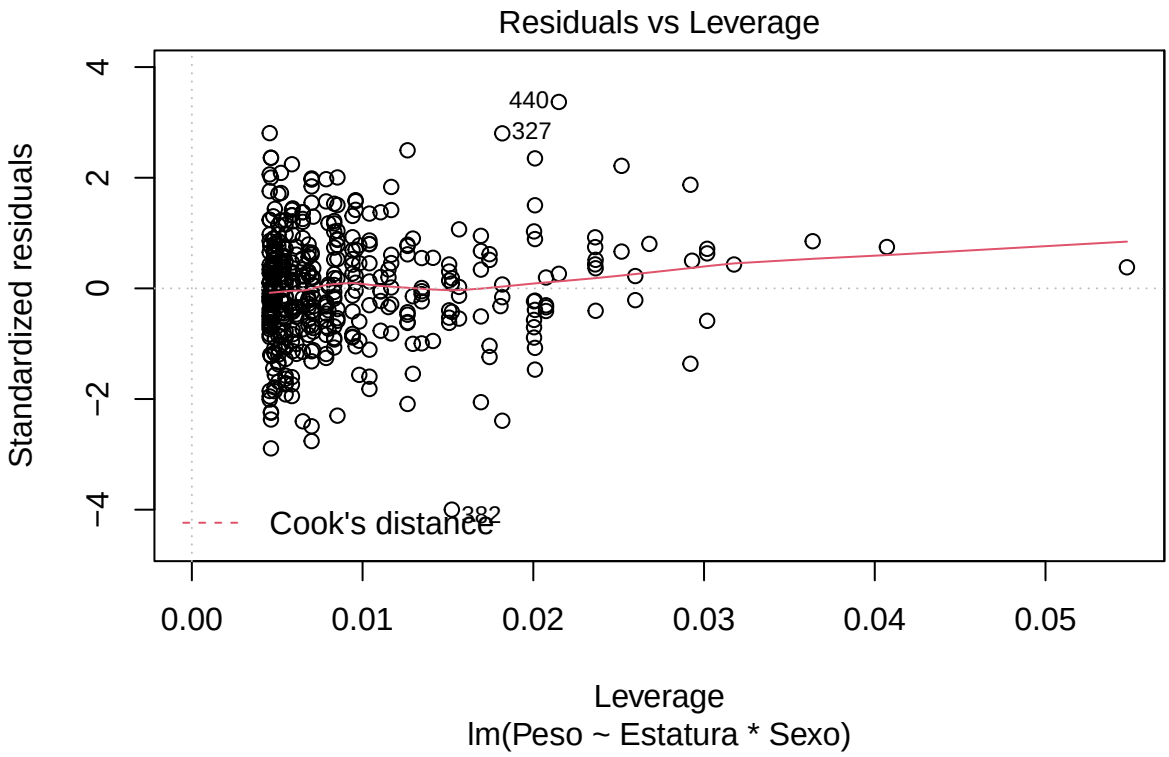



```
plot(Modelo3)
```


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


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Peso ~ Estatura * 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(Peso ~ Estatura * Sexo)



¿Cuáles son las diferencias y similitudes de estos gráficos con respecto a los que ya habías analizado?

Estos gráficos son muy similares a los que ya habíamos obtenido, pues obtenemos los mismos resultados.

Estos gráficos, ¿cambian en algo las conclusiones que ya habías obtenido?

Estos gráficos no cambian en nada los resultados que ya habíamos obtenido, pues se observa el mismo comportamiento: * Hay independencia para todos los modelos. * Hay normalidad en todos los modelos. * La media de los residuos sigue siendo cero en todos los modelos. * La varianza solo se observa constante para el primer modelo en el conjunto de los hombres.

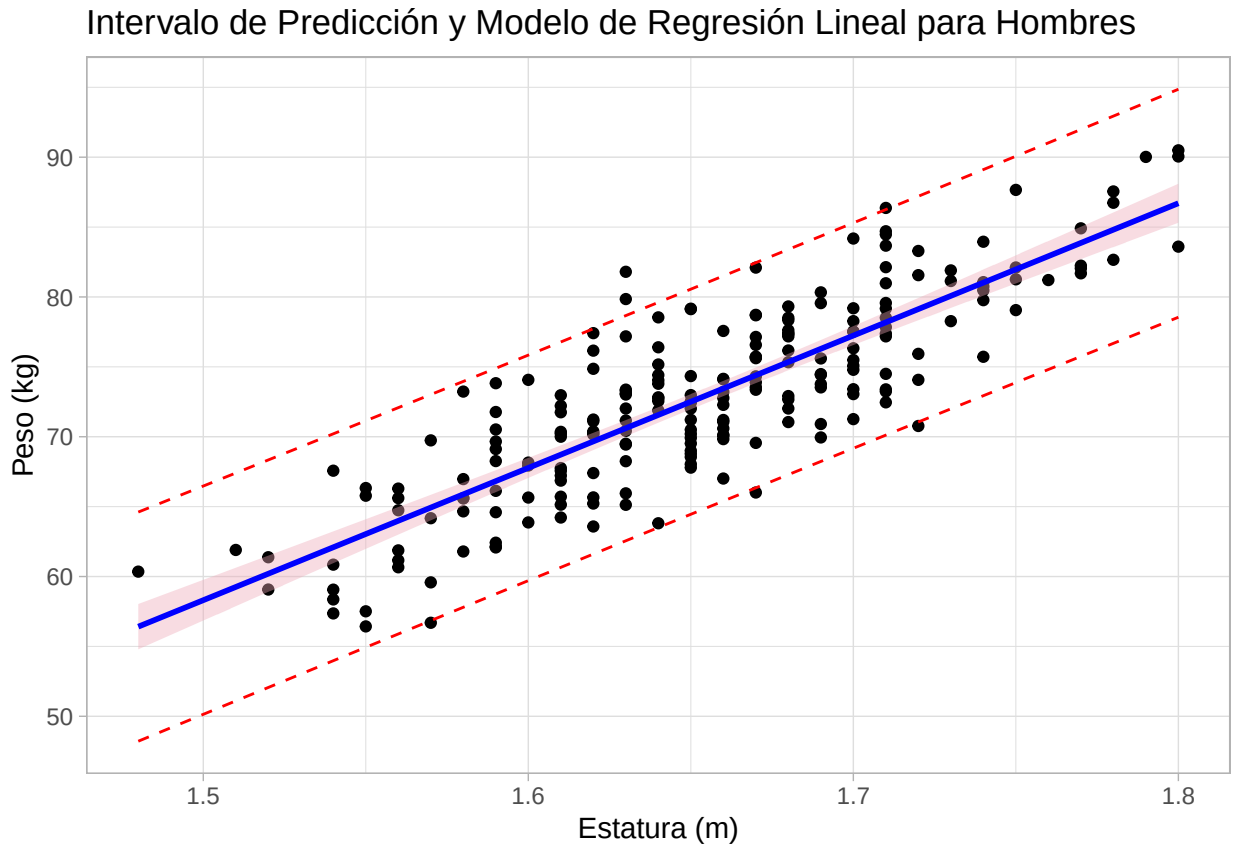
Intervalos de confianza

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 mejor modelo seleccionado.

```
suppressWarnings({
A  = Modelo1H
Ip = predict(object=A, interval="prediction", level=0.97)
datos1 = cbind(MH, Ip)

library(ggplot2)
ggplot(datos1, aes(x= MH$Estatura, y= MH$Peso)) +
  geom_point() +
```

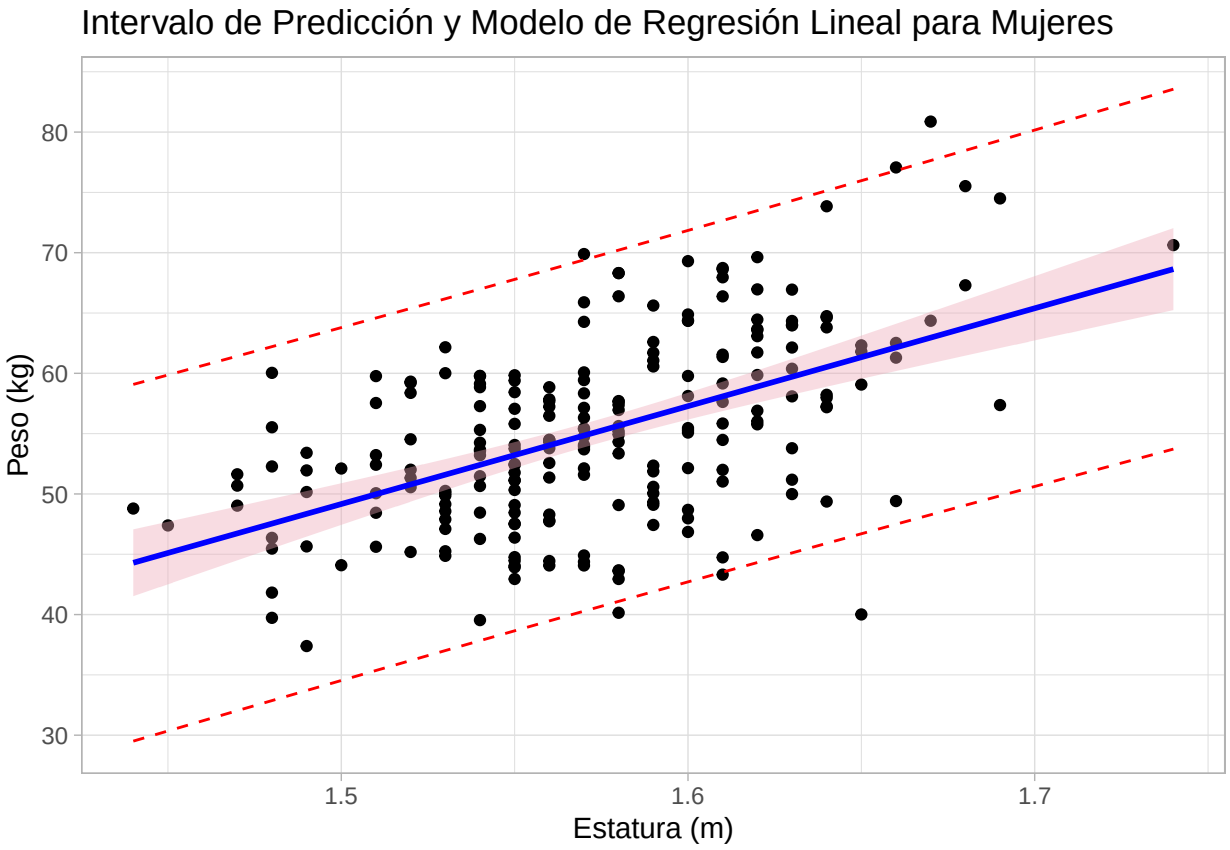
```
  geom_line(aes(y=lwr), color="red", linetype="dashed") +
  geom_line(aes(y=upr), color="red", linetype="dashed") +
  geom_smooth(method=lm, formula=y~x, se=TRUE, level=0.97, col="blue", fill="pink2") +
  labs(
     title = "Intervalo de Predicción y Modelo de Regresión Lineal para Hombres",
    x = "Estatura (m)",
    y = "Peso (kg)"
  ) +
  theme_light()
})
```




```
suppressWarnings({
B = Modelo1M
Ip = predict(object=B, interval="prediction", level=0.97)
datos2 = cbind(MM, Ip)

ggplot(datos2, aes(x= MM$Estatura, y= MM$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") +
  labs(
    title = "Intervalo de Predicción y Modelo de Regresión Lineal para Mujeres",
    x = "Estatura (m)",
    y = "Peso (kg)"
```

```
  ) +
  theme_light()
})
```



Interpretación

Las líneas rojas punteadas representan los Intervalos de Predicción. Estos intervalos nos dan un rango en el cual se espera que se encuentre la próxima observación realizada con un 97% de confianza o seguridad.

La línea azul representa el modelo de regresión lineal. Es decir, es donde el modelo realiza la predicción media de la próxima observación. Es decir, predice el peso promedio de la siguiente observación realizada.

El área sombreada roja representa el intervalo de Confianza para la regresión lineal realizada por el modelo. En otras palabras, es donde se espera que se encuentra el verdadero valor de la próxima observación realizada con un 97% de confianza.