Actividad 12. Regresión Lineal

Oscar Gutierrez

2024-08-30

Análisis descriptivo

```
M = read.csv('Estatura-peso_HyM.csv')
head(M)
```

```
##   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
```

Matriz de correlación

```
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
```

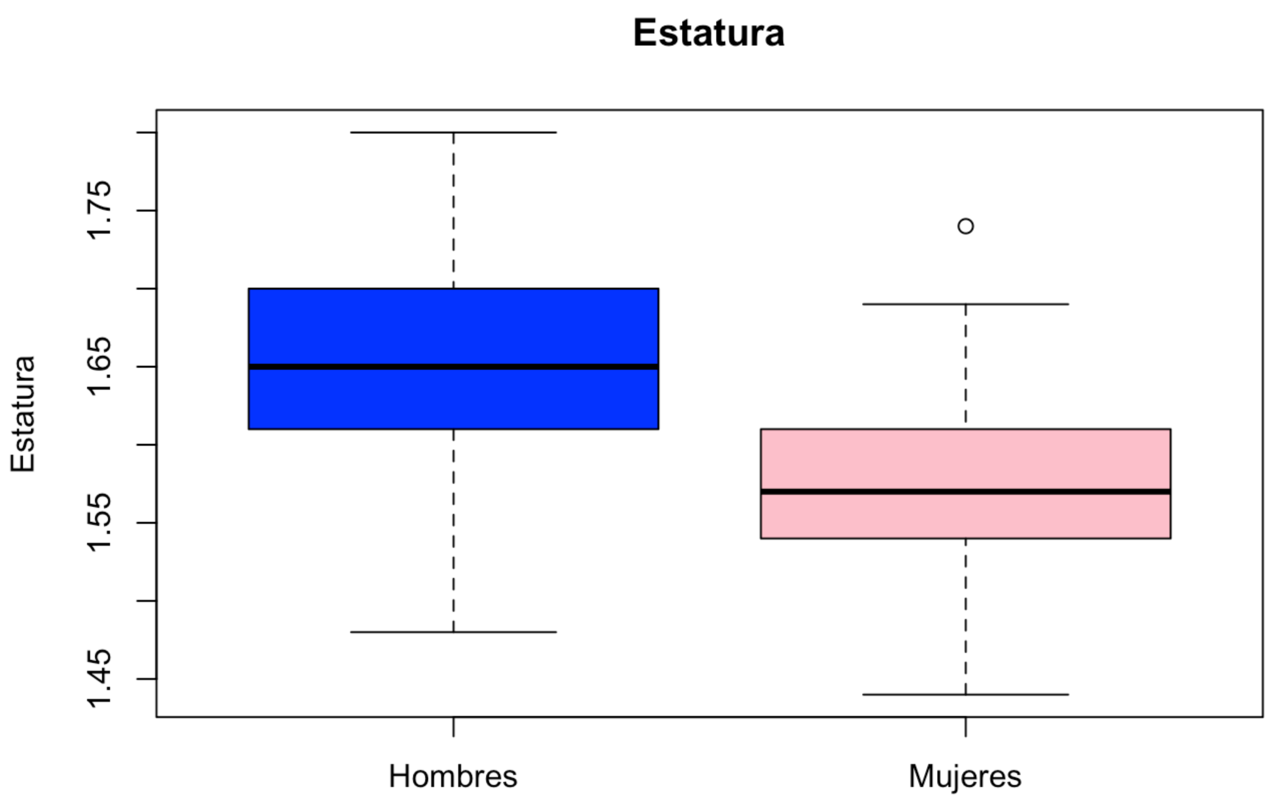
Se observa que hay una correlación de 0.85 entre el peso y estatura de los hombres y una correlación de 0.52 entre el peso y estatura de las mujeres.

Medidas relevantes

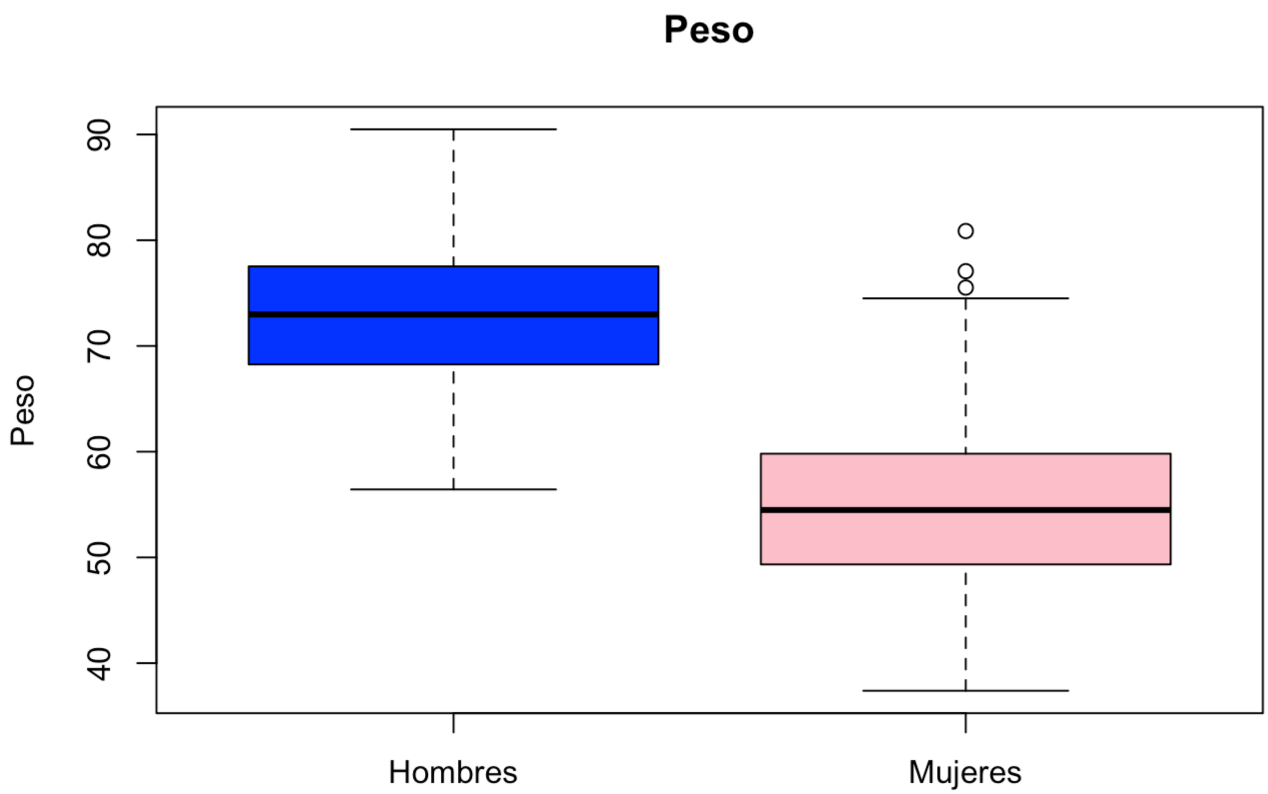
```
n=4 #número de variables
d=matrix(NA,ncol=7,nrow=n)
for(i in 1:n){
  d[i,]<-c(as.numeric(summary(M1[,i])),sd(M1[ ,i]))
}
m=as.data.frame(d)

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
```

```
boxplot(M$Estatura~M$Sexo, ylab="Estatura", xlab="", col=c("blue","pink"), names=c("Homb
res", "Mujeres"), main="Estatura")
```




```
boxplot(M$Peso~M$Sexo, ylab="Peso",xlab="", names=c("Hombres", "Mujeres"), col=c("blu
e","pink"), main="Peso")
```



Modelos separados

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

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

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

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

Hipótesis $H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$

Con $\alpha = 0.03$

Hombres

```
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
```

El 71% de la varianza es explicada por el modelo, el otro 29% es debido a los errores. También se encuentra que el peso y el intercept son significativos.

Mujeres

```
summary(Modelo1M)
```

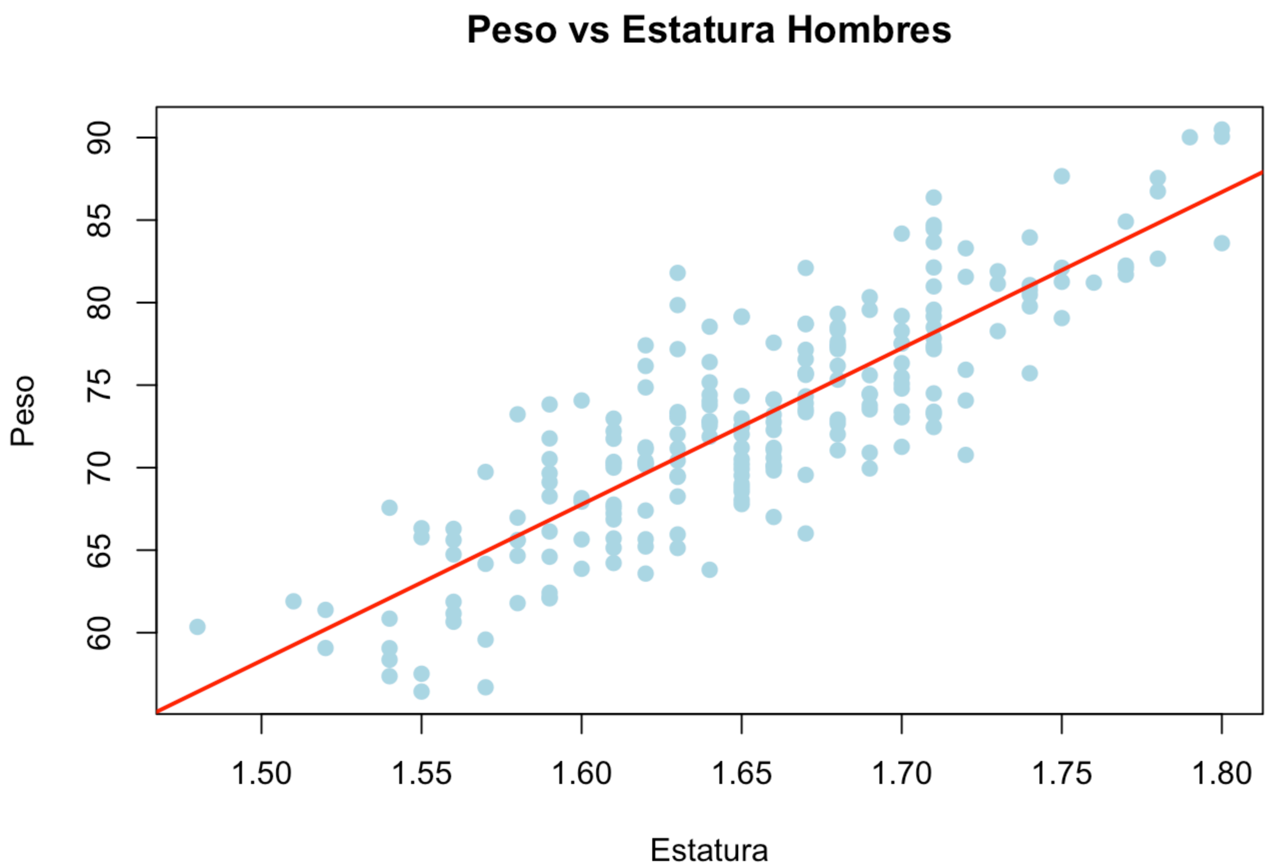
```
##
## Call:
## lm(formula = Peso ~ Estatura, data = MM)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3256  -4.1942   0.4004   4.2724  17.9114
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -72.560     14.041  -5.168 5.34e-07 ***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

El 27.51% de la varianza es explicada por el modelo, el otro 72.49% es debido a los errores. Esto quiere decir que el modelo no es capaz de explicar la mayoría de la variación. Se encuentra que el peso y el intercept son significativos.

Gráficas

Hombres

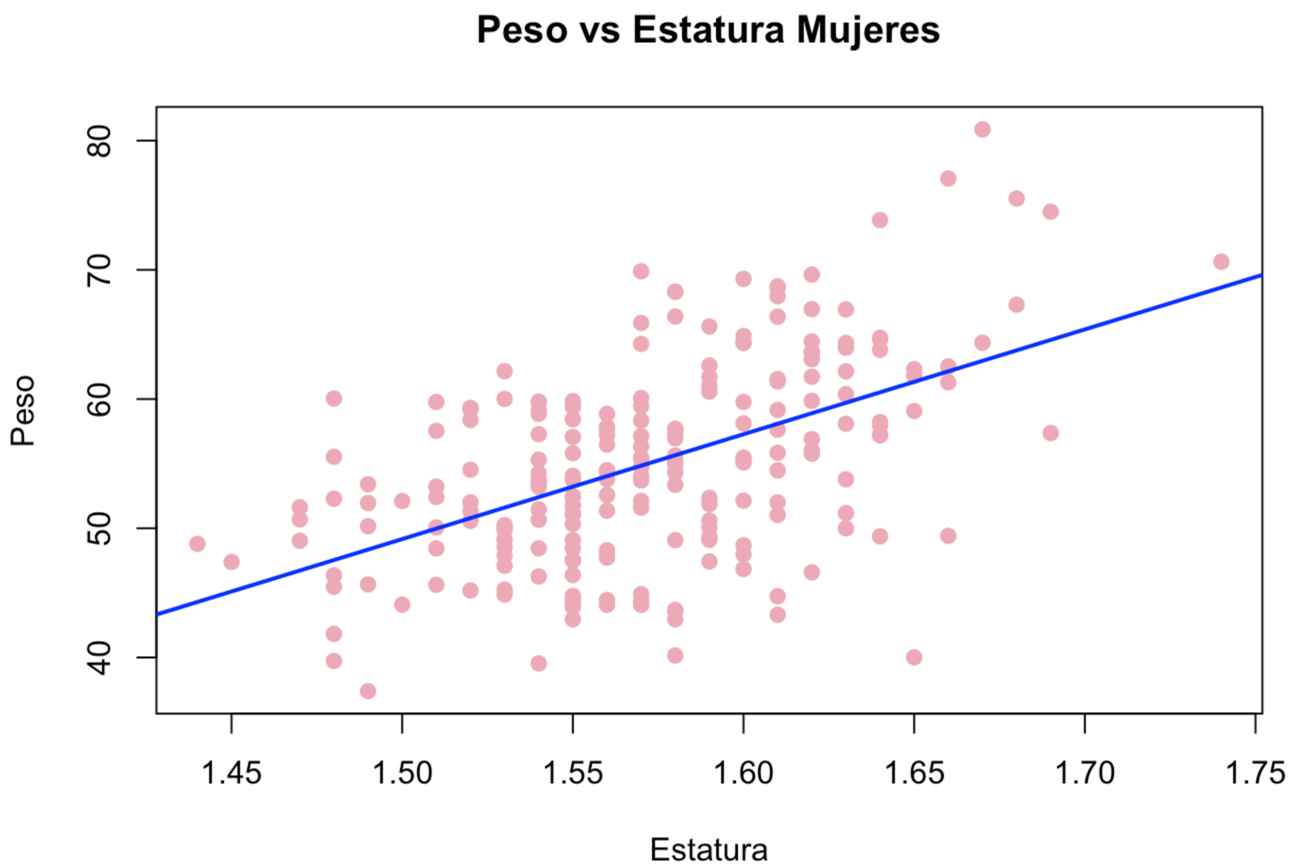
```
plot(MH$Estatura, MH$Peso, main="Peso vs Estatura Hombres",
     xlab="Estatura", ylab="Peso", pch=19, col="lightblue")
abline(Modelo1H, col="red", lwd=2)
```



Esta recta aproxima correctamente los datos.

Mujeres

```
plot(MM$Estatura, MM$Peso, main="Peso vs Estatura Mujeres",
     xlab="Estatura", ylab="Peso", pch=19, col="pink2")
abline(Modelo1M, col="blue", lwd=2)
```



La recta obtenida no predice correctamente los valores para peso puesto que en estos datos hay más variabilidad.

Un modelo con los sexos juntos

```
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
```

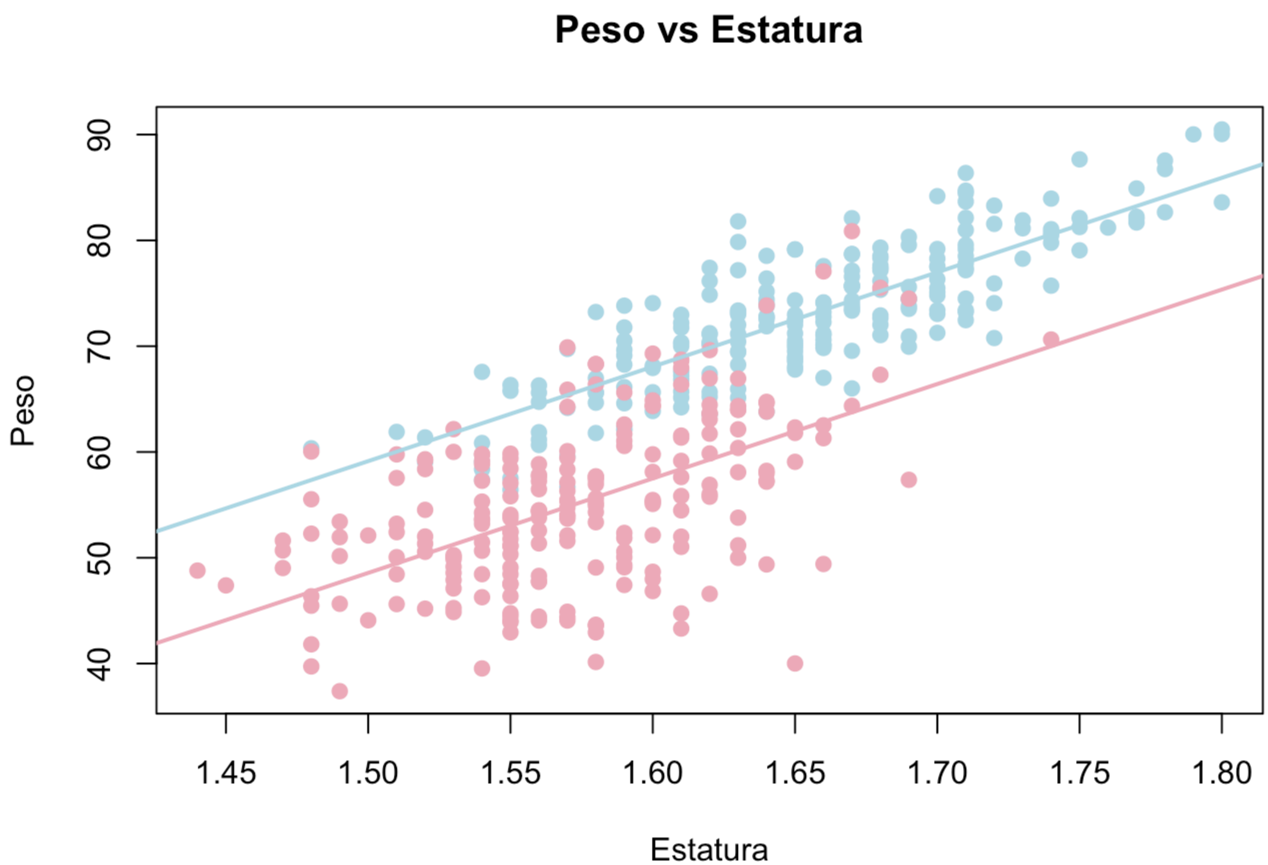
```
summary(Modelo2)
```

```
##
## 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
```

El intercept, la estatura y el sexo son significativos. El modelo logra explicar un 78% de la variación, el resto es debido a errores.

Grafico

```
b0 = Modelo2$coefficients[1]
b1 = Modelo2$coefficients[2]
b2 = Modelo2$coefficients[3]

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

colores = c('lightblue', 'pink2')

plot(M$Estatura, M$Peso, main="Peso vs Estatura",
     xlab="Estatura", ylab="Peso", pch=19, col=colores[factor(M$Sexo)])

x = seq(min(M$Estatura)*0.9, max(M$Estatura)*1.1, 0.01)
lines(x, Ym(x), col = "pink2", lwd = 2)
lines(x, Yh(x), col = "lightblue", lwd = 2)
```



A pesar de que el modelo junto da un coeficiente de determinación más alto, tiene la desventaja que asigna el mismo coeficiente para la estatura, solamente cambia el intercept dependiendo del Sexo.

β_0 indica el peso cuando la altura es 0, que a pesar que esto es imposible, en este contexto ayuda a ver que tan altas son las personas en general, por eso cambia dependiendo del sexo puesto que en general, los hombres son más altos que las mujeres.

β_1 indica cuánto se espera que cambie el peso por cada incremento de una unidad en la estatura.

Un modelo 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
```

```
A = summary(Modelo3)
A
```

```
##
## 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
```

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

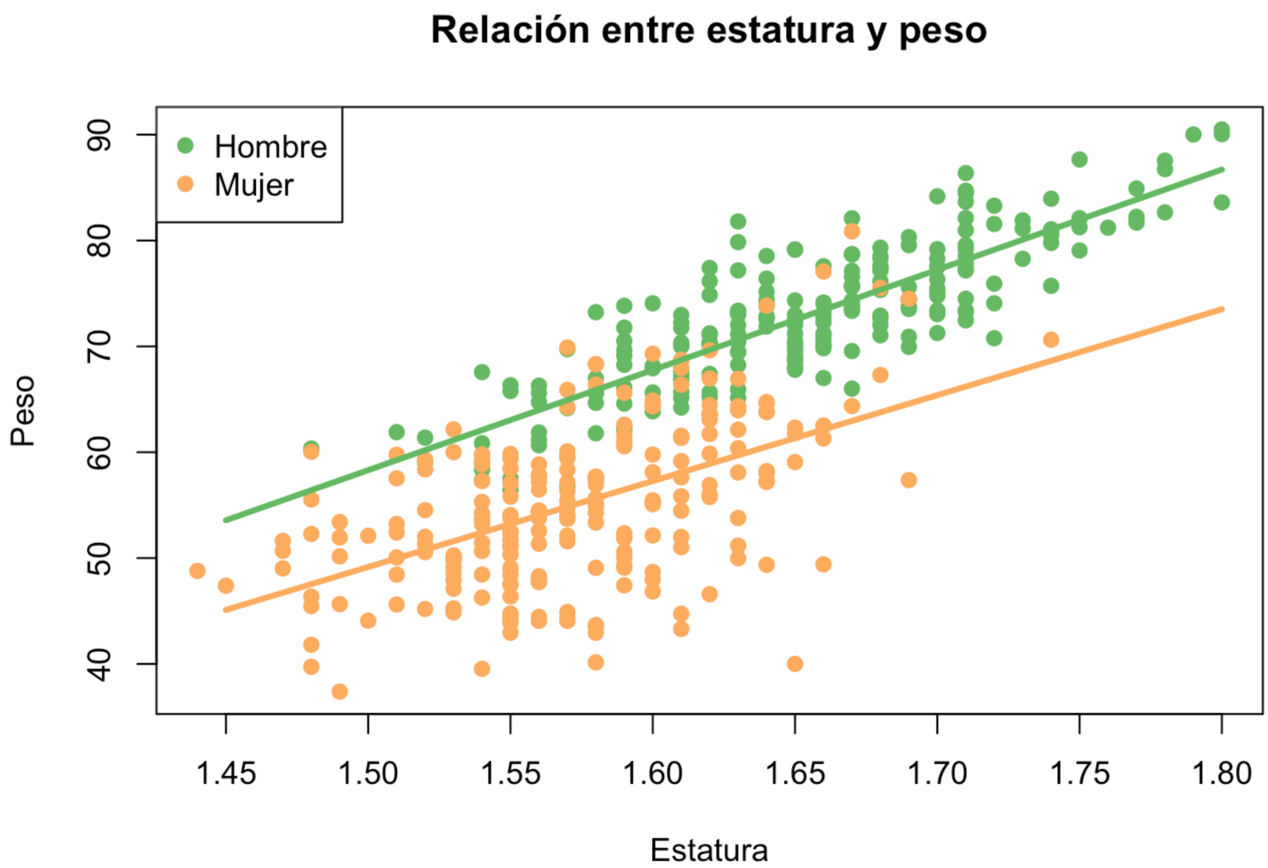
Con $\alpha = 0.03$

La variable dummy SexoM tiene valor 1 cuando es mujer y 0 cuando es hombre.

En este modelo se obtiene que el Sexo no es significativo ya que el valor p es mayor a α. El modelo y el resto de las variables son significativos.

El porcentaje de variación explicada por el modelo es de aproximadamente 78%

```
b0_A <- A$coefficients[1]
b1_A <- A$coefficients[2]
b2_A <- A$coefficients[3]
b3_A <- A$coefficients[4]

Yh <- function(x) { b0_A + b1_A * x }
Ym <- function(x) { b0_A + b2_A + (b1_A + b3_A) * x }

colores <- c("#66BD63","#FDAE61" )
plot(M$Estatura, M$Peso, col=colores[factor(M$Sexo)], pch=19, ylab="Peso", xlab="Estatur
a", main="Relación entre estatura y peso")

x <- seq(1.45, 1.80, 0.01)
lines(x, Yh(x), col="#66BD63", lwd=3)  # Line for males
lines(x, Ym(x), col="#FDAE61", lwd=3)  # Line for females

legend("topleft", legend=c("Hombre", "Mujer"), pch=19, col=c("#66BD63", "#FDAE61"))
```



Conclusiones

Se realizaron 3 modelos, uno donde se dividieron hombres y mujeres, otro donde se mantuvieron juntos y otro considerando la interacción entre sexo y estatura. Los coeficientes obtenidos del modelo de regresión son similares entre el primer y el tercer modelo. Por otro lado, el segundo modelo tiene menos 'flexibilidad' puesto que no puede asignar un valor diferente para la pendiente dependiendo del sexo, solamente puede ajustar el intercept, en cambio, el primer y el tercer modelo pueden ajustar estos dos valores dependiendo del sexo.

El intercept (β_0) es la intersección en el eje y cuando la estatura es igual a 0, en este contexto no tiene un significado como tal..

El coeficiente que acompaña a la estatura (β_1) explica cuanto cambia el peso por cada incremento de una unidad en la estatura, el coeficiente que acompaña a la variable dummy SexoM cambia el intercept dependiendo del sexo, y el coeficiente que acompaña la interacción modifica la pendiente dependiendo del Sexo. El mejor modelo para realizar un análisis de este contexto es el tercero, puesto que permite identificar todos los elementos que conforman el problema ya que toma en consideración la estatura, el sexo y la interacción entre ellos. A diferencia del primero y el segundo, en el primero se separan los grupos y a pesar de que es posible hacer el análisis de esta manera es mucho más claro en el tercer modelo, en el segundo modelo se tiene un modelo menos flexible porque no se puede ajustar la pendiente dependiendo del sexo, lo cual sí es posible en el tecer modelo.

Normalidad de residuos

H_0 : Los residuos se distribuyen normalmente H_1 : Los residuos no se distribuyen normalmente

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

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

```r
ad.test(Modelo1M$residuals)
```

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

```r
ad.test(Modelo2$residuals)
```

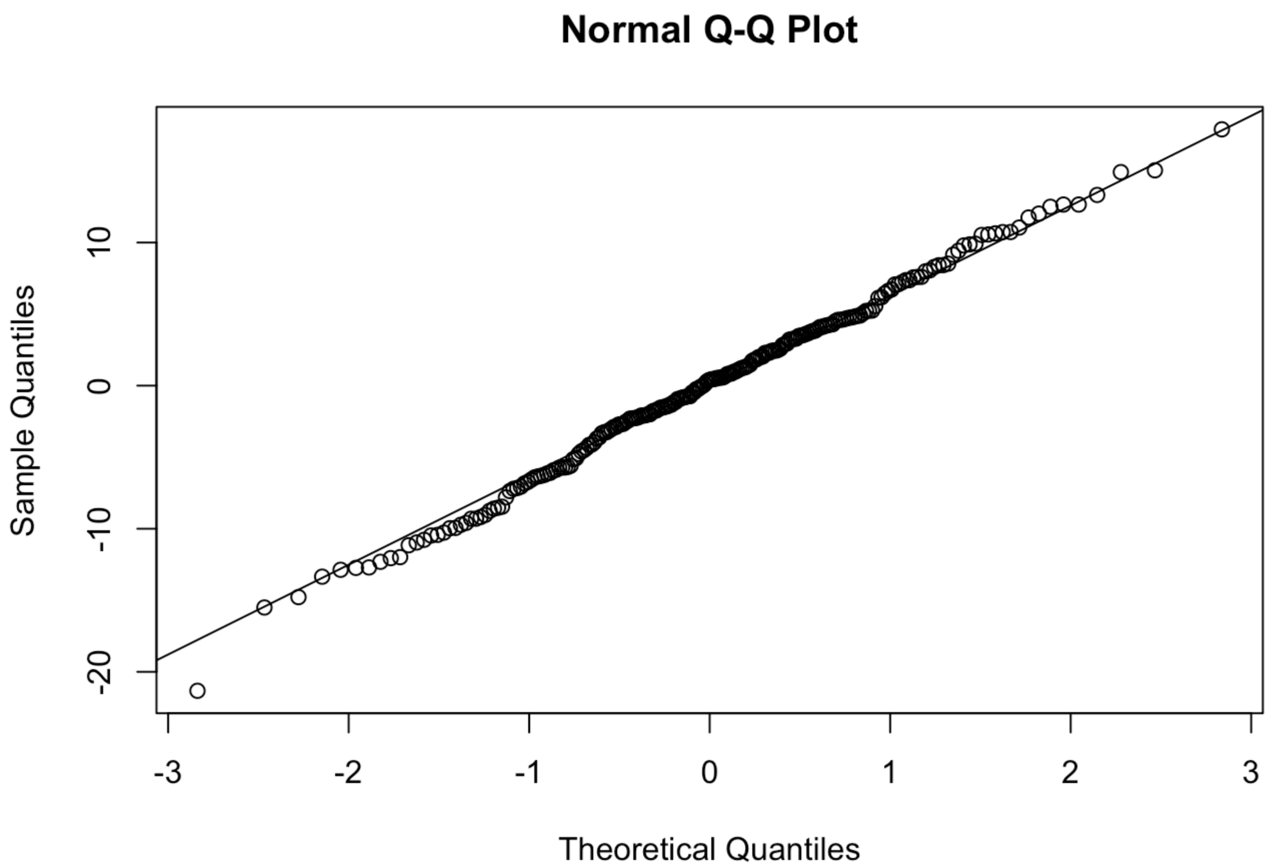
```
##
##  Anderson-Darling normality test
##
## data:  Modelo2$residuals
## A = 0.79651, p-value = 0.03879
```

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

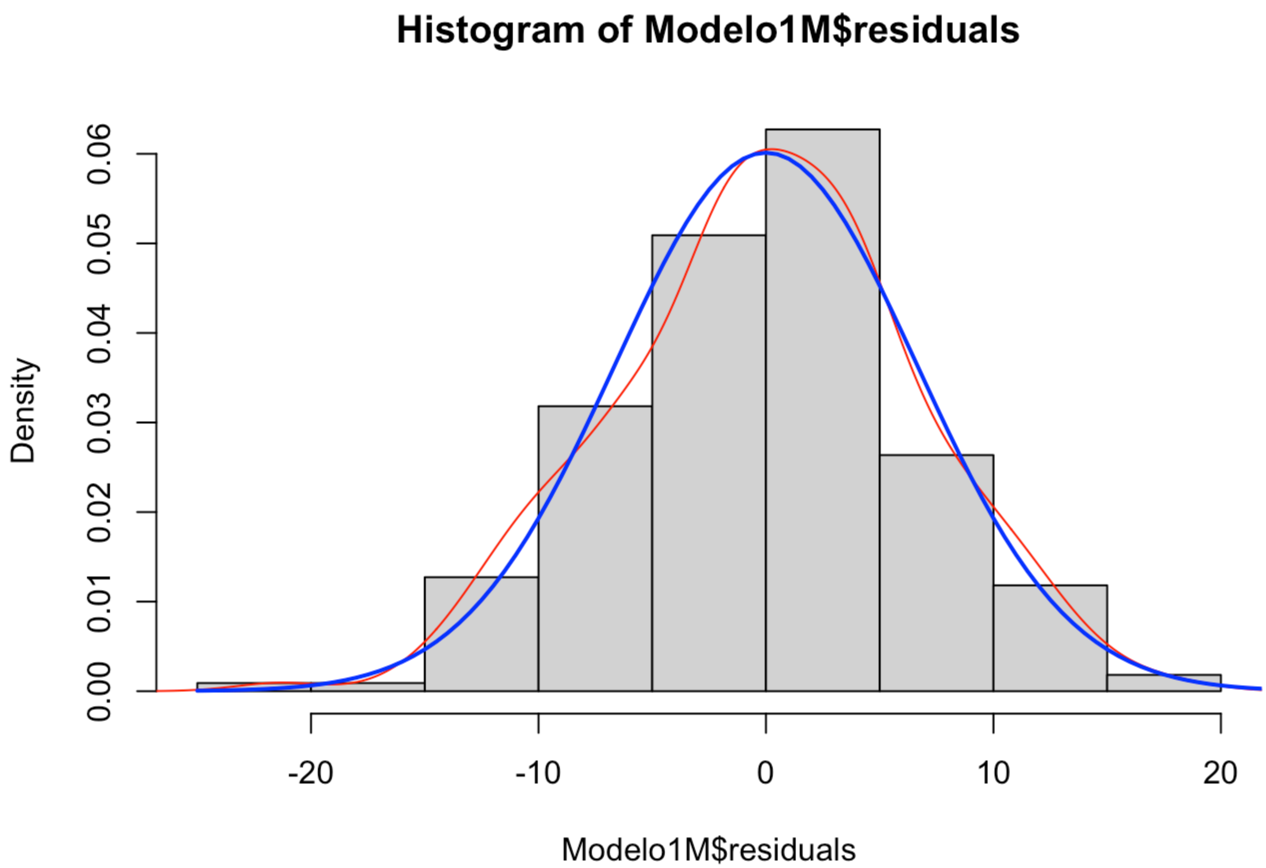
```
##
##  Anderson-Darling normality test
##
## data:  Modelo3$residuals
## A = 0.8138, p-value = 0.03516
```

No se tiene evidencia para rechazar H_0 en ninguno de los modelos ya que el valor p es mayor a $\alpha = 0.3$

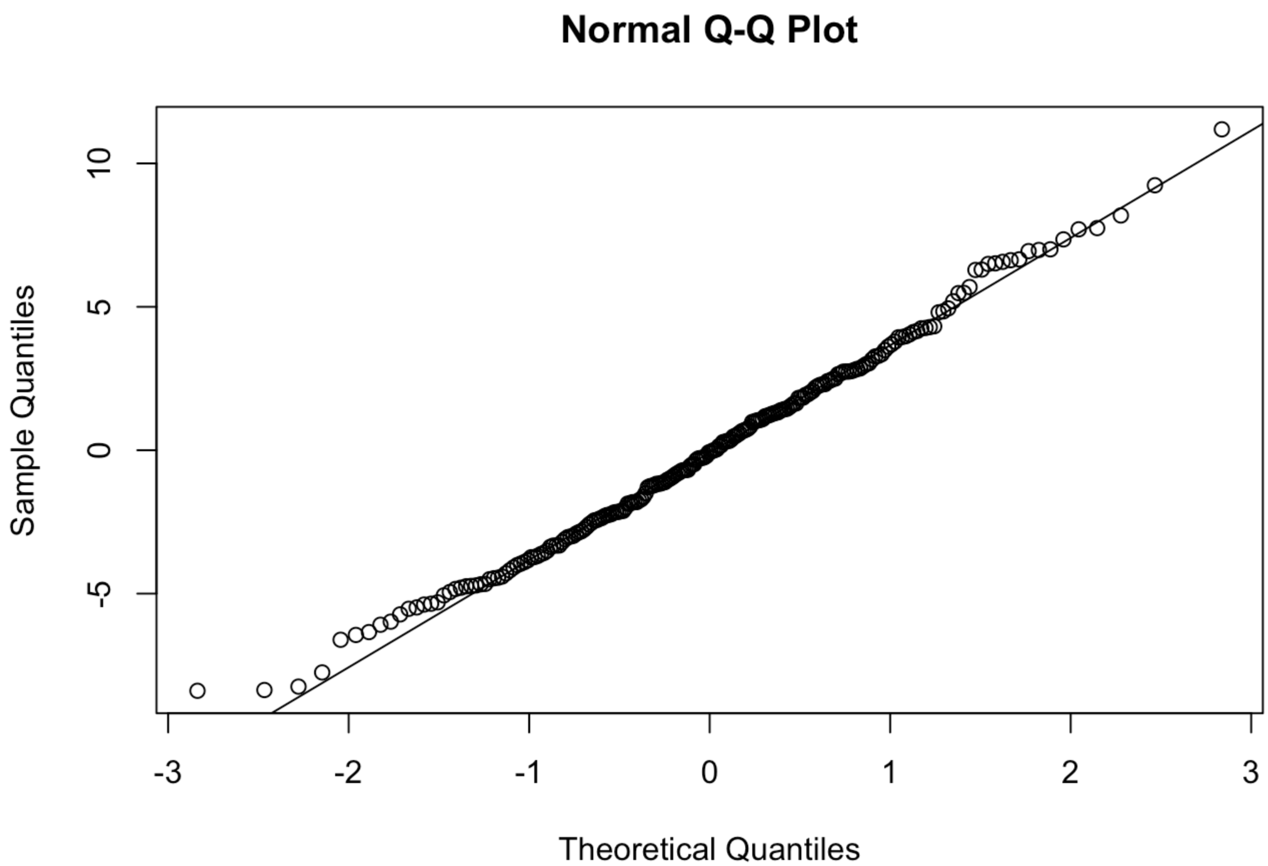
```r
# Modelo 1 Mujeres
qqnorm(Modelo1M$residuals)
qqline(Modelo1M$residuals)
```



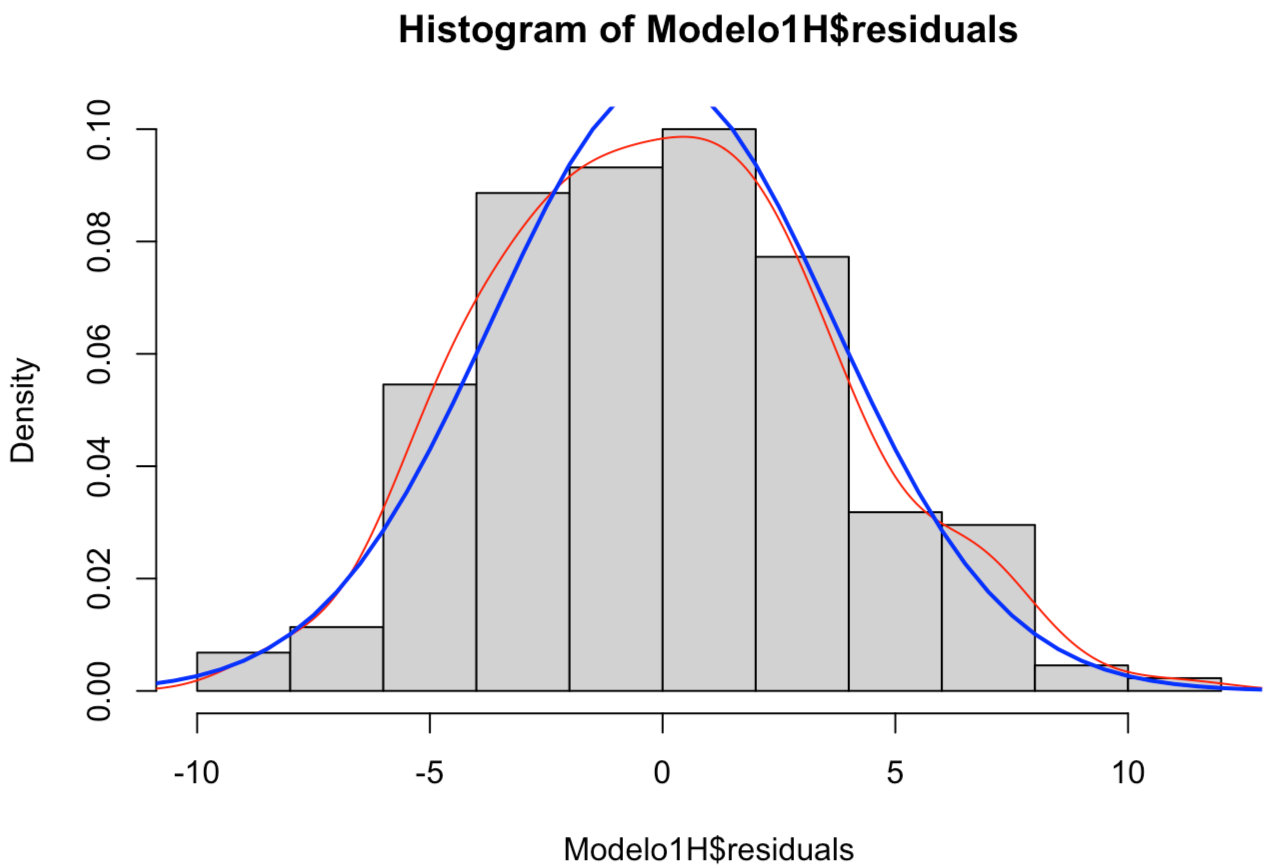

```
hist(Modelo1M$residuals,freq=FALSE)
lines(density(Modelo1M$residual),col="red")
curve(dnorm(x,mean=mean(Modelo1M$residuals),sd=sd(Modelo1M$residuals)), from=-25, to=25,
add=TRUE, col="blue",lwd=2)
```



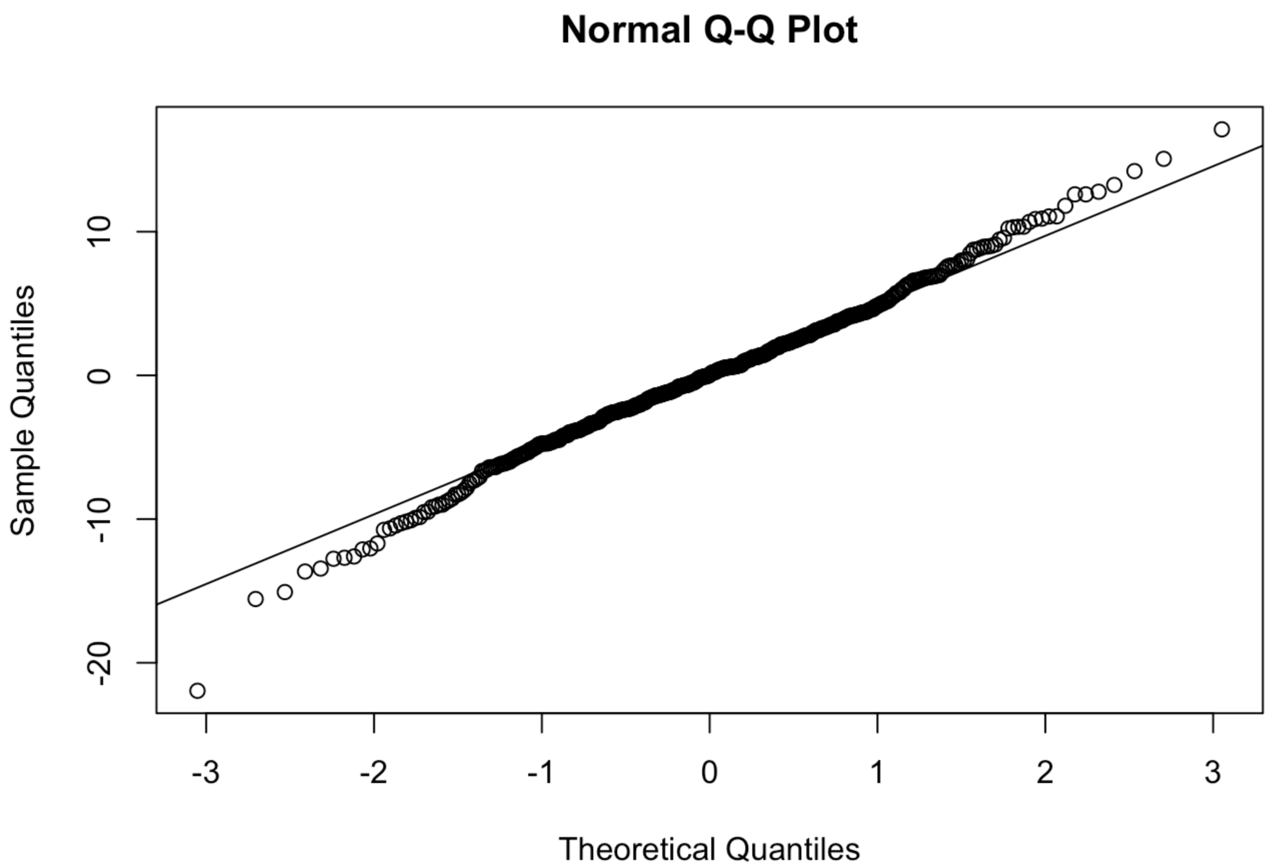

```
# Modelo 1 Hombres
qqnorm(Modelo1H$residuals)
qqline(Modelo1H$residuals)
```



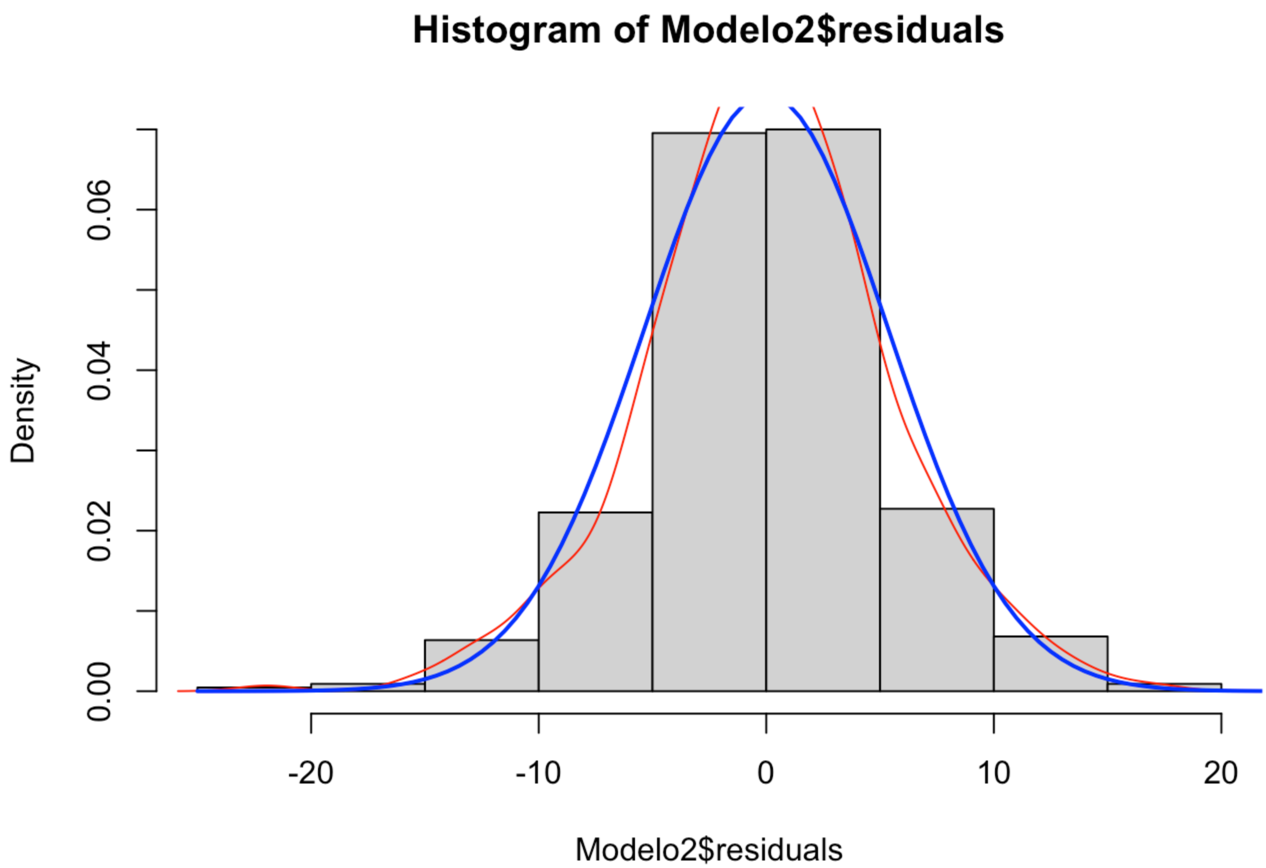

```
hist(Modelo1H$residuals,freq=FALSE)
lines(density(Modelo1H$residual),col="red")
curve(dnorm(x,mean=mean(Modelo1H$residuals),sd=sd(Modelo1H$residuals)), from=-25, to=25,
add=TRUE, col="blue",lwd=2)
```



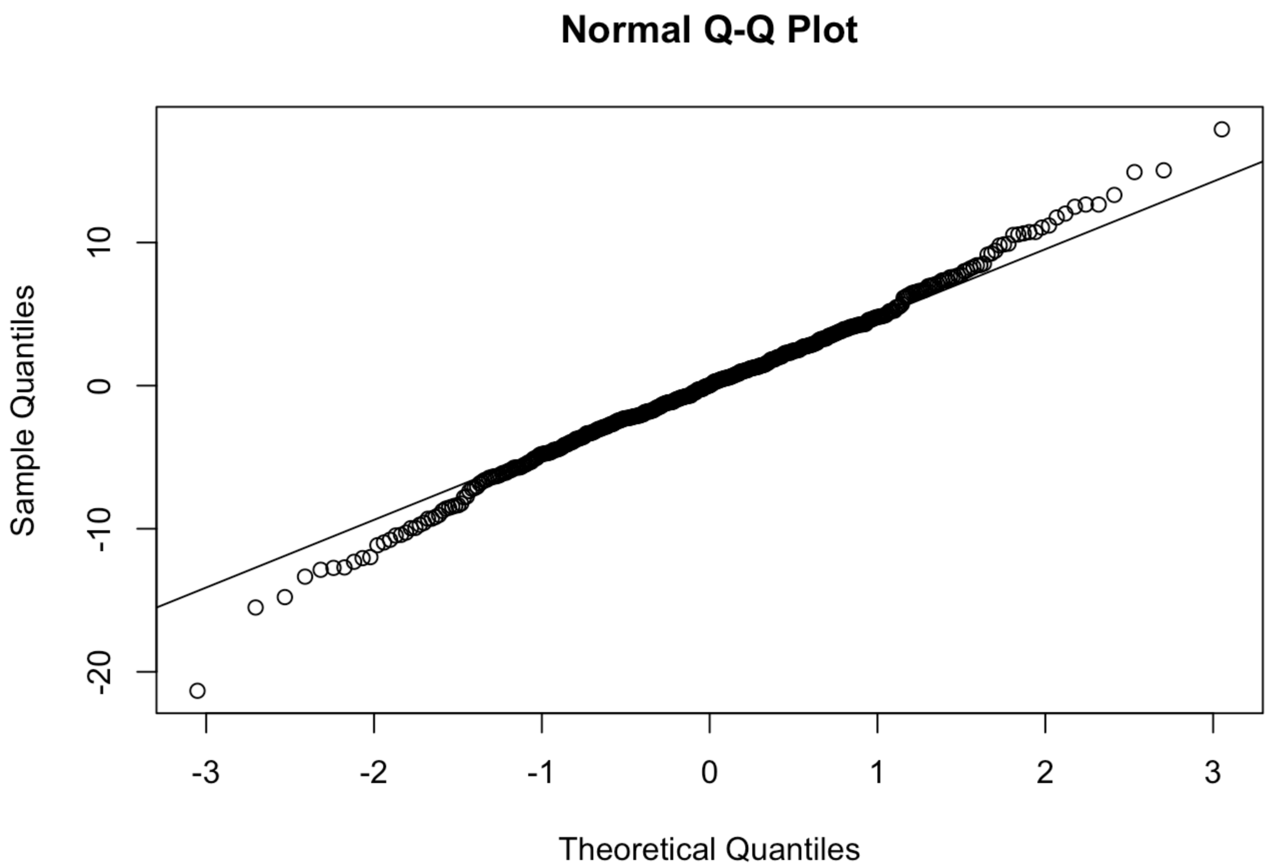

```
# Modelo 2
qqnorm(Modelo2$residuals)
qqline(Modelo2$residuals)
```



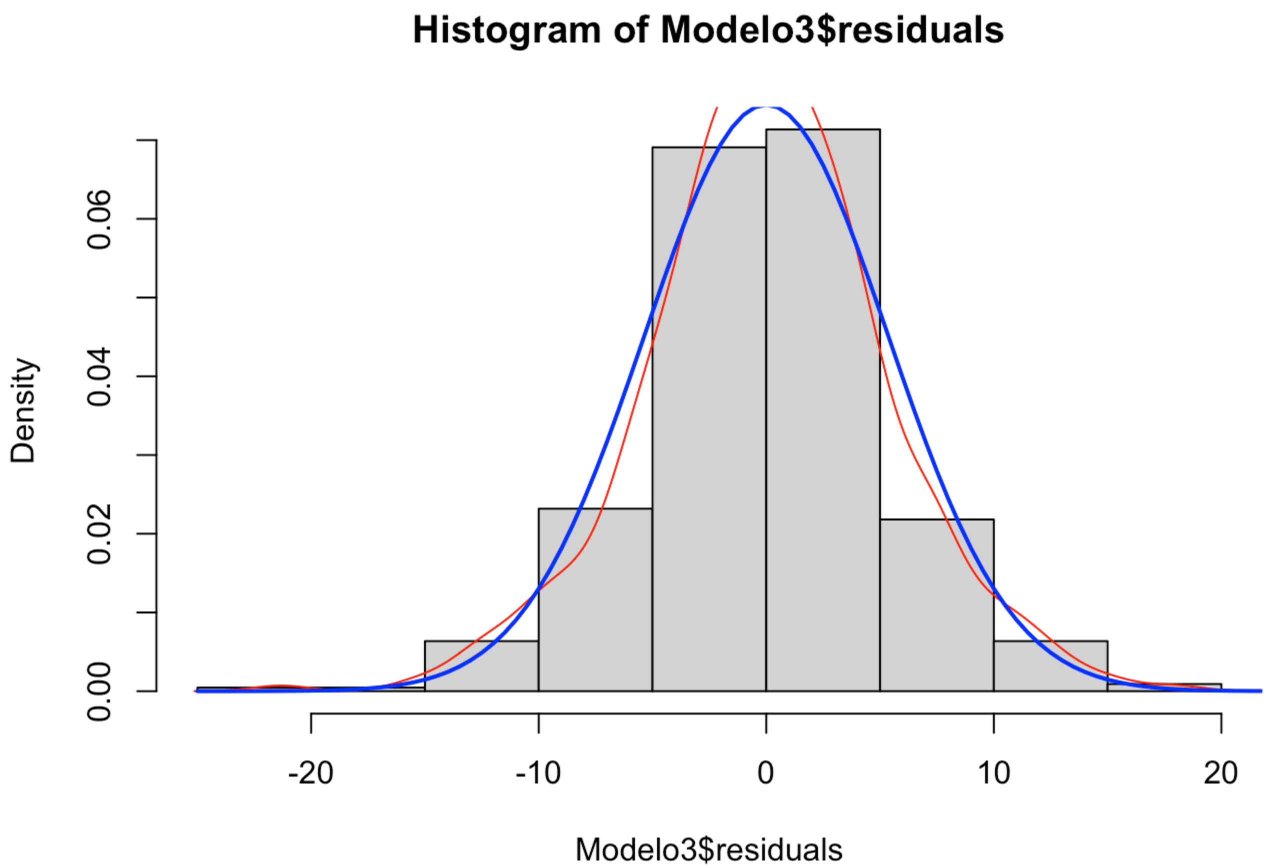

```
hist(Modelo2$residuals,freq=FALSE)
lines(density(Modelo2$residual),col="red")
curve(dnorm(x,mean=mean(Modelo2$residuals),sd=sd(Modelo2$residuals)), from=-25, to=25, a
dd=TRUE, col="blue",lwd=2)
```




```
# Modelo 3
qqnorm(Modelo3$residuals)
qqline(Modelo3$residuals)
```




```
hist(Modelo3$residuals,freq=FALSE)
lines(density(Modelo3$residual),col="red")
curve(dnorm(x,mean=mean(Modelo3$residuals),sd=sd(Modelo3$residuals)), from=-25, to=25, a
dd=TRUE, col="blue",lwd=2)
```



Comprobar media = 0

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

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

```
##
##  One Sample t-test
##
## data:  Modelo1M$residuals
## t = 7.2202e-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
## 3.22974e-16
```

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

```
##
##  One Sample t-test
##
## data:  Modelo1H$residuals
## t = 9.3819e-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
## 2.321375e-16
```

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

```
##
##  One Sample t-test
##
## data:  Modelo2$residuals
## t = 6.941e-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
## 1.776357e-16
```

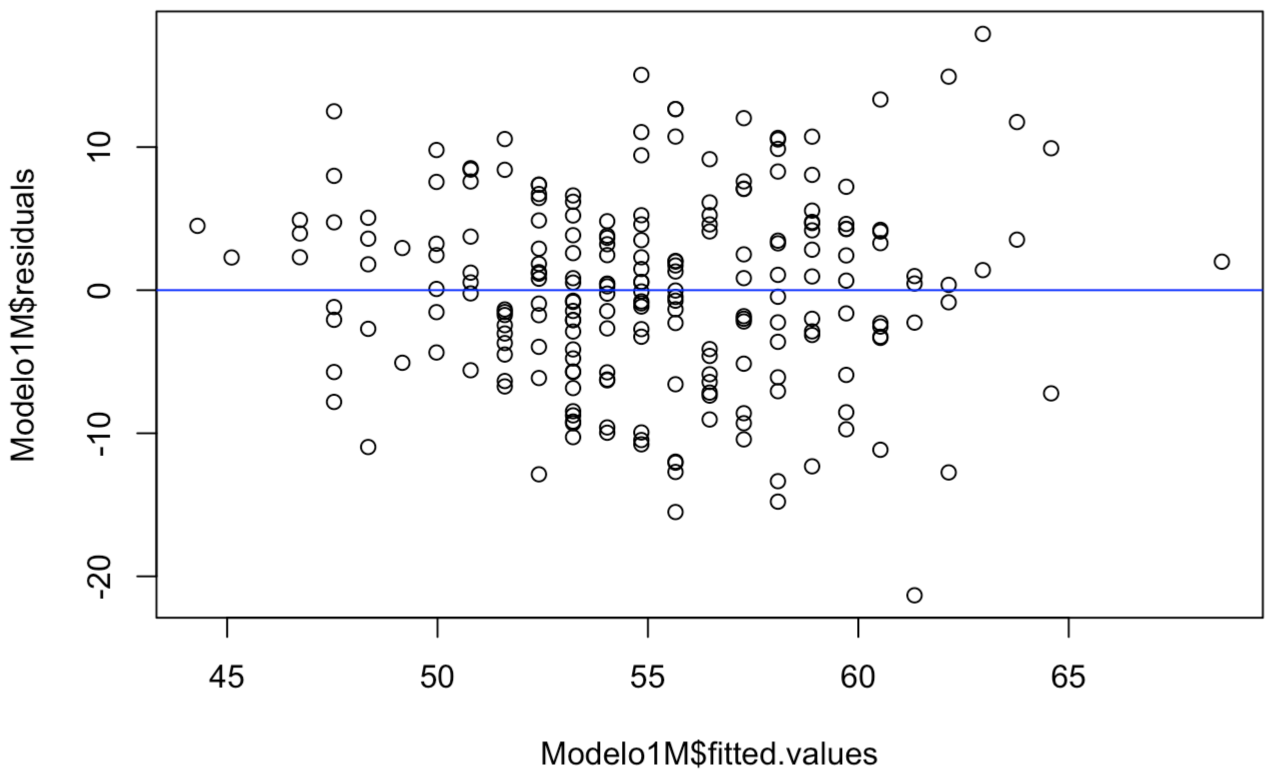
```
t.test(Modelo3$residuals)
```

```
##
##  One Sample t-test
##
## data:  Modelo3$residuals
## t = -3.1626e-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
## -8.074349e-17
```

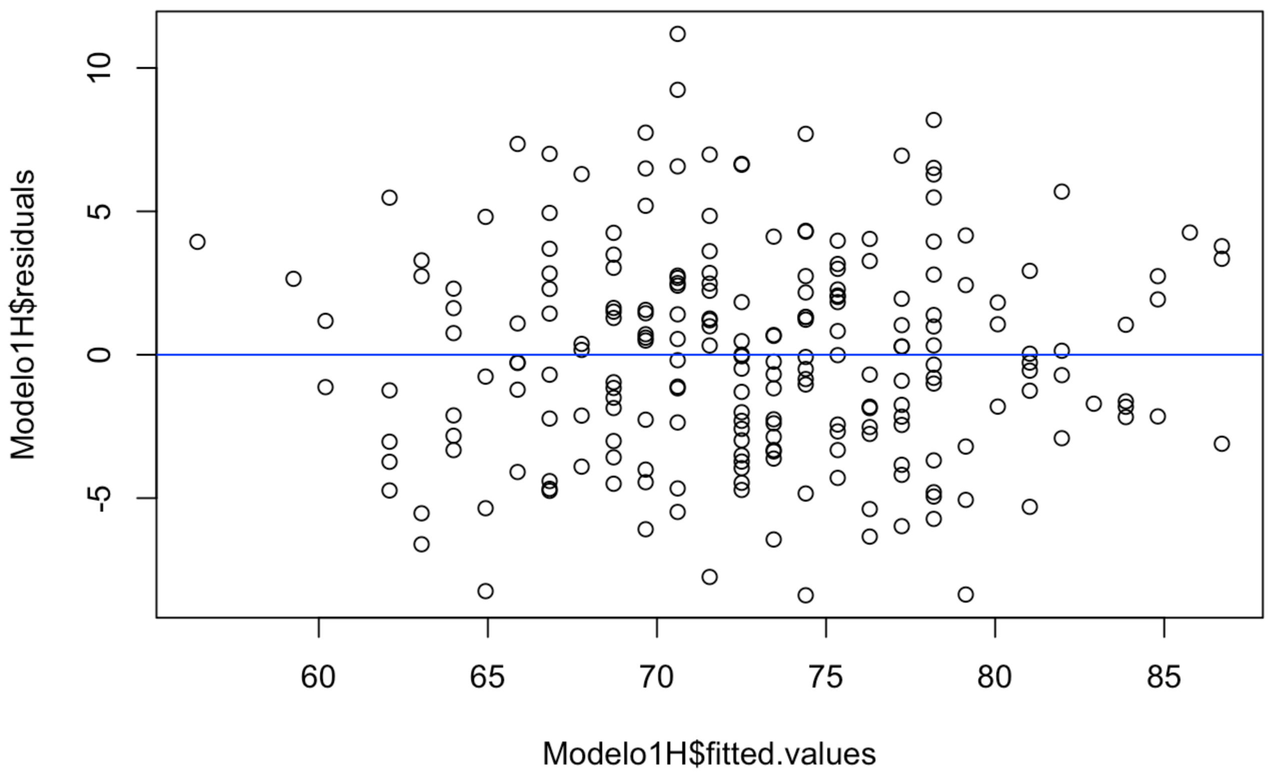
No se tiene evidencia para rechazar la hipótesis nula en ningun modelo ya que el valor p es mayor a $\alpha = 0.3$

Homocedasticidad e independencia

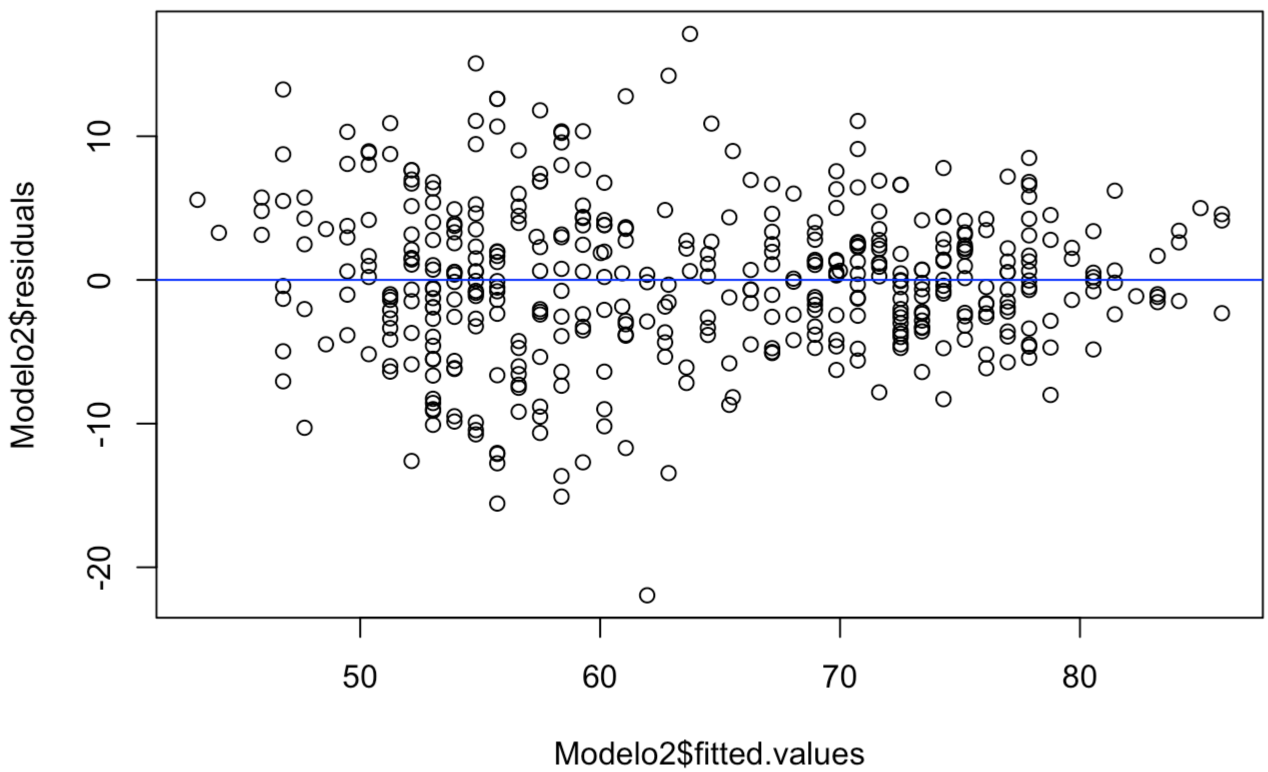
```
# Modelo 1M
plot(Modelo1M$fitted.values, Modelo1M$residuals)
abline(h=0, col= 'blue')
```



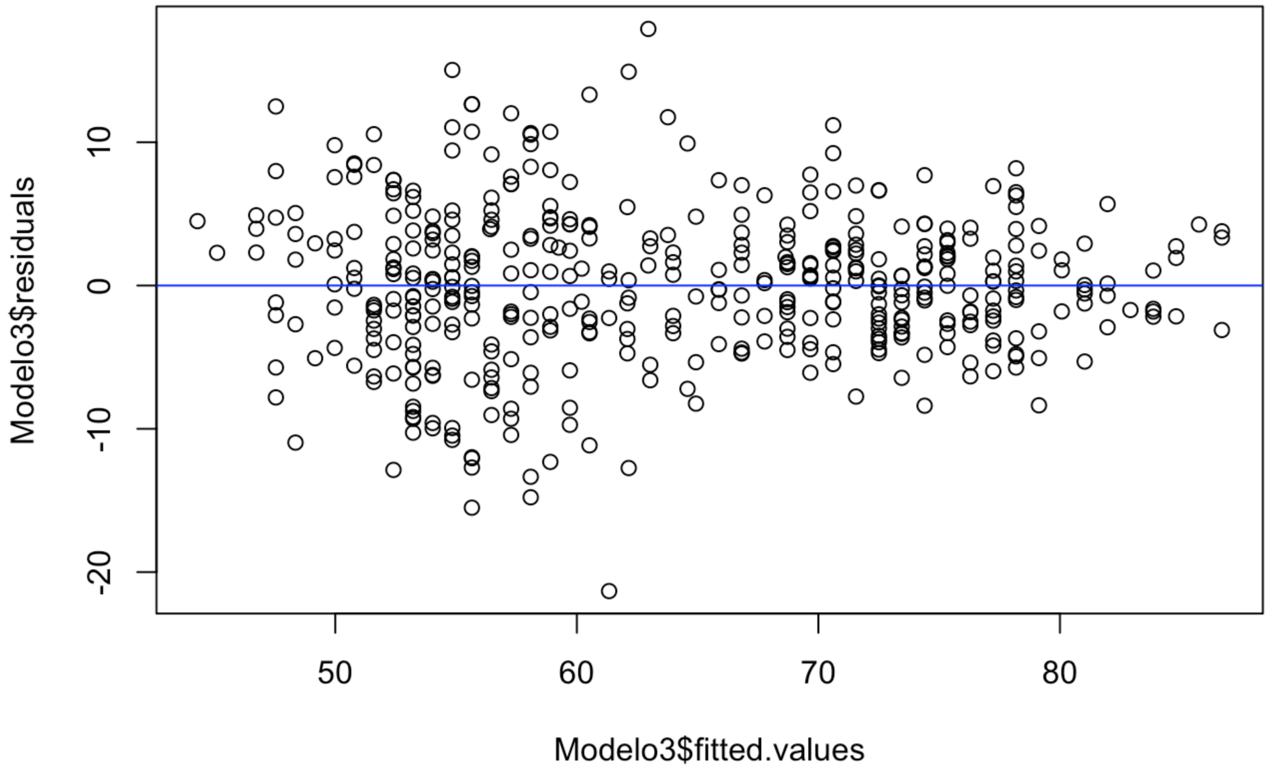

```
# Modelo 1H
plot(Modelo1H$fitted.values, Modelo1H$residuals)
abline(h=0, col= 'blue')
```




```
# Modelo 2
plot(Modelo2$fitted.values, Modelo2$residuals)
abline(h=0, col= 'blue')
```




```
# Modelo 3
plot(Modelo3$fitted.values, Modelo3$residuals)
abline(h=0, col= 'blue')
```



Los errores se ven aleatorios, no parecen seguir alguna tendencia en particular.

Prueba de independencia

H_0: Los errores no están autocorrelacionados. H_1: Los errores están autocorrelacionados.

```
library(lmtest)
```

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

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

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

```
dwtest(Modelo1M)
```

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

```
bgtest(Modelo1M)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  Modelo1M
## LM test = 1.4655, df = 1, p-value = 0.2261
```

```
dwtest(Modelo1H)
```

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

```
bgtest(Modelo1H)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  Modelo1H
## LM test = 0.20778, df = 1, p-value = 0.6485
```

```
dwtest(Modelo2)
```

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

```
bgtest(Modelo2)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  Modelo2
## LM test = 1.3595, df = 1, p-value = 0.2436
```

```
dwtest(Modelo3)
```

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

```
bgtest(Modelo3)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  Modelo3
## LM test = 1.3453, df = 1, p-value = 0.2461
```

Ninguna de las pruebas rechaza la hipótesis nula, los errores no están autocorrelacionados debido a que los valores p son mayores a α.

Prueba de homocedasticidad

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

```
gqtest(Modelo1M)
```

```
##
##  Goldfeld-Quandt test
##
## data:  Modelo1M
## GQ = 1.4265, df1 = 108, df2 = 108, p-value = 0.03313
## alternative hypothesis: variance increases from segment 1 to 2
```

```
gqtest(Modelo1H)
```

```
##
##  Goldfeld-Quandt test
##
## data:  Modelo1H
## GQ = 0.84148, df1 = 108, df2 = 108, p-value = 0.8144
## alternative hypothesis: variance increases from segment 1 to 2
```

```
gqtest(Modelo2)
```

```
##
##  Goldfeld-Quandt test
##
## data:  Modelo2
## GQ = 3.2684, df1 = 217, df2 = 217, p-value < 2.2e-16
## alternative hypothesis: variance increases from segment 1 to 2
```

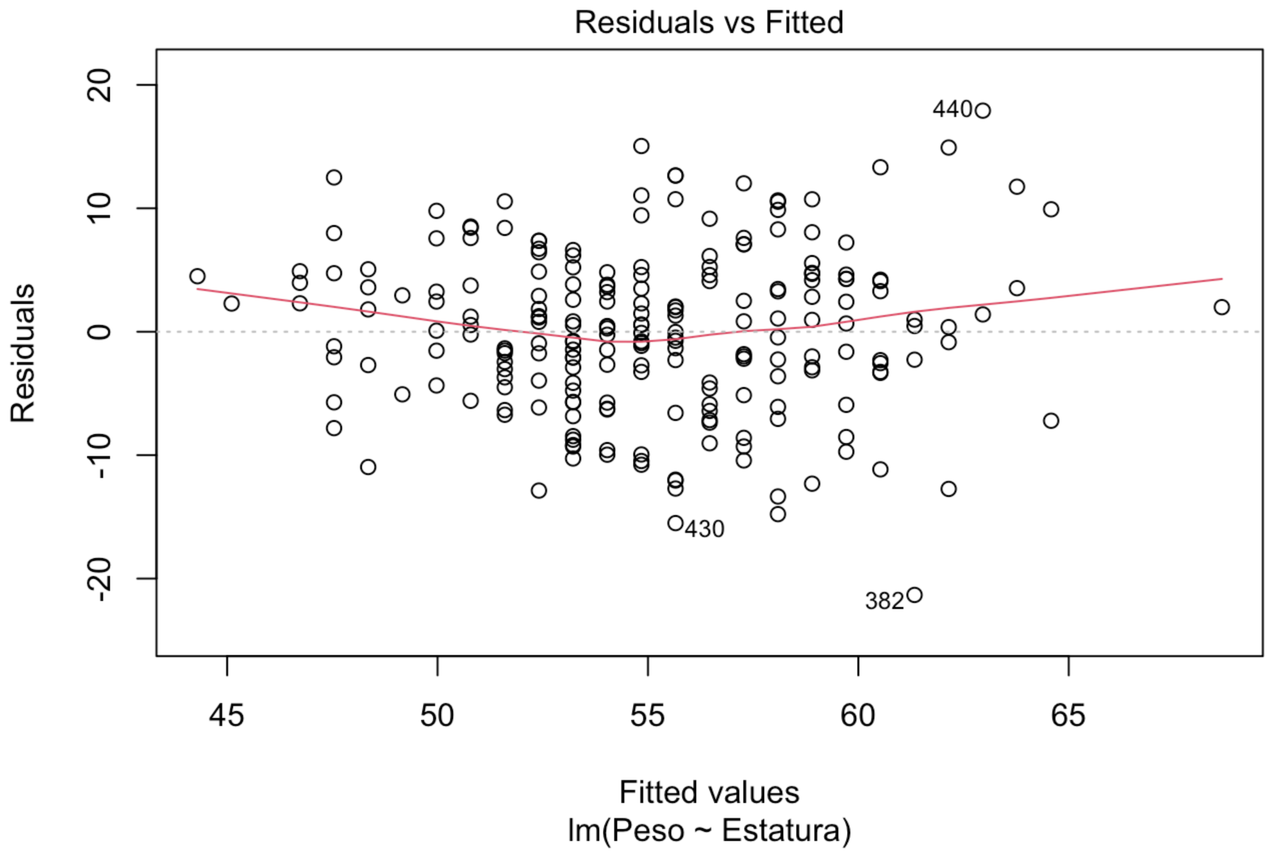
```
gqtest(Modelo3)
```

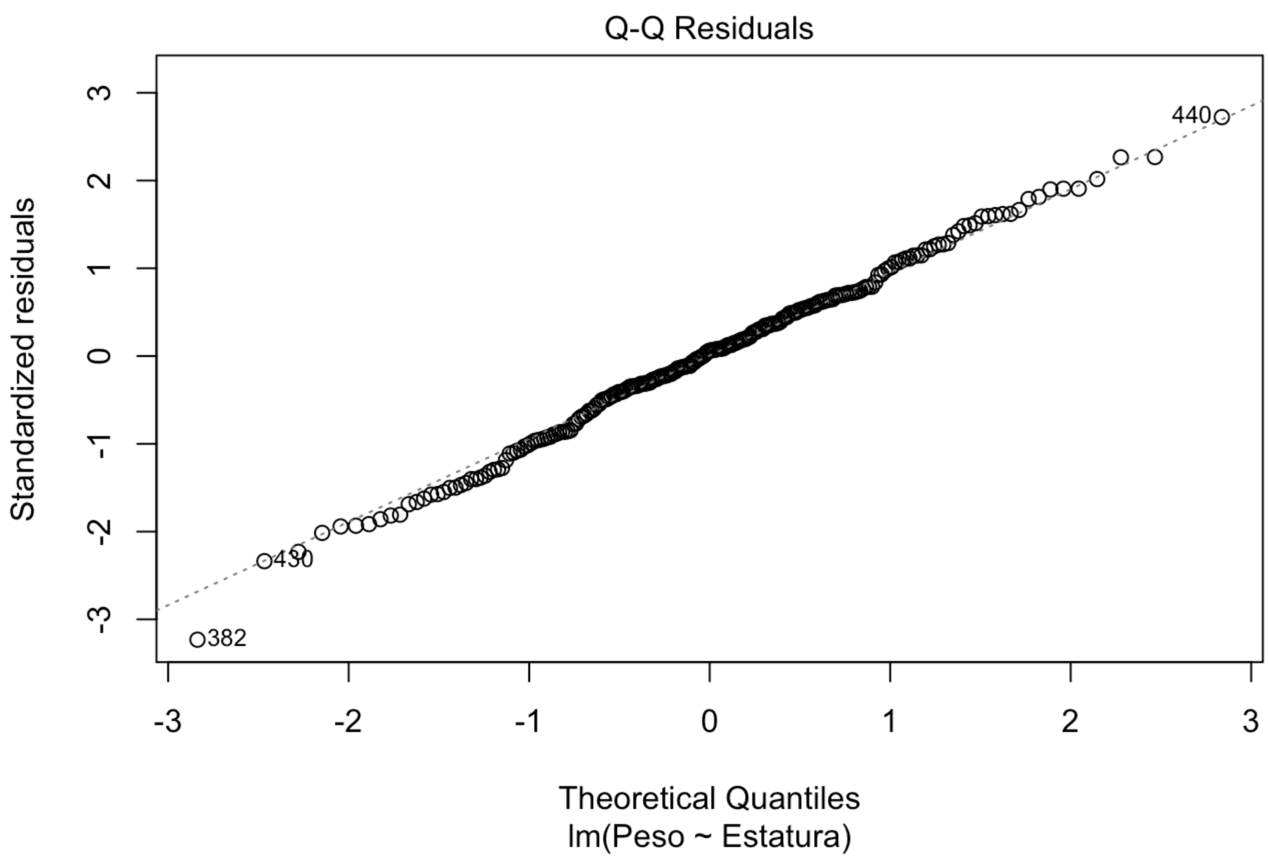
```
##
##  Goldfeld-Quandt test
##
## data:  Modelo3
## GQ = 3.2684, df1 = 216, df2 = 216, p-value < 2.2e-16
## alternative hypothesis: variance increases from segment 1 to 2
```

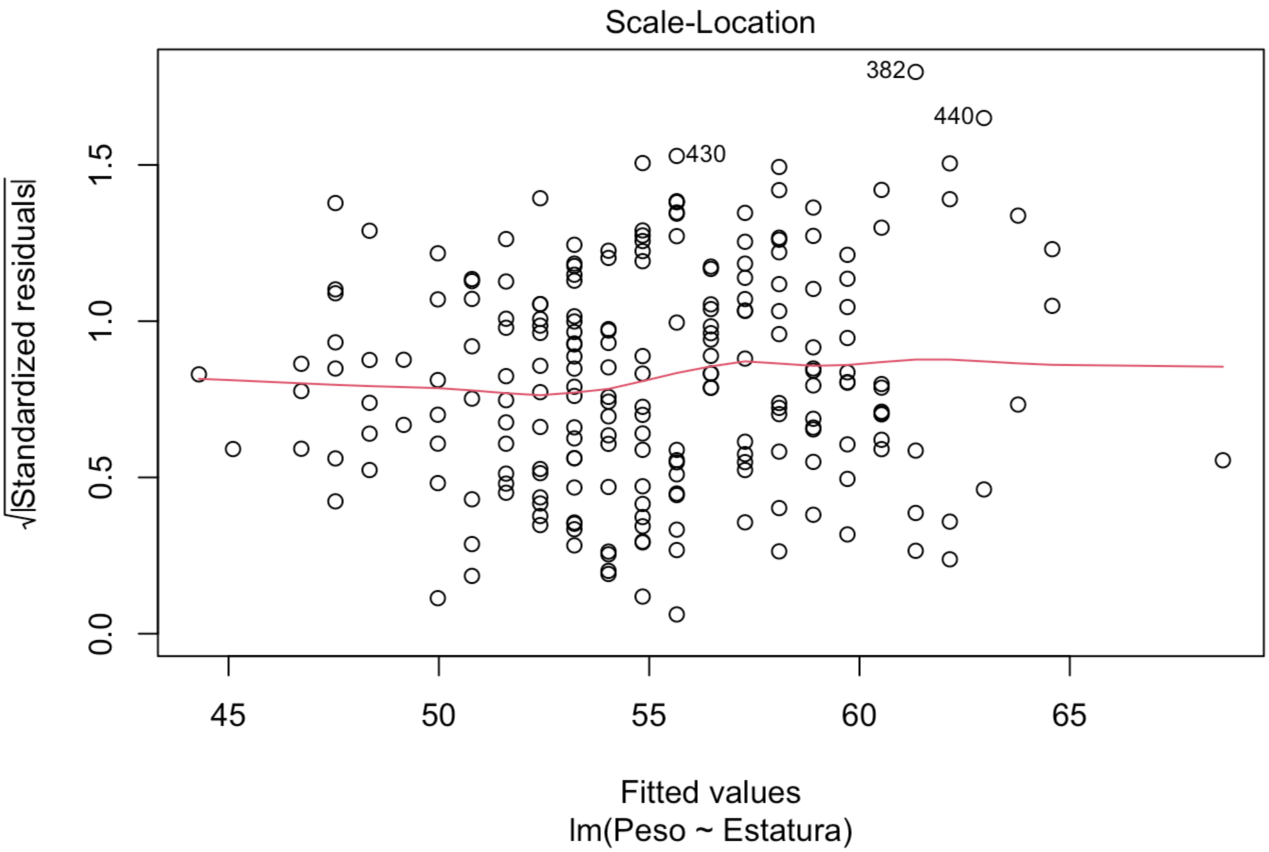
Solamente se mantiene la hipótesis nula en el caso del Modelo1, en el resto se rechaza la hipótesis nula, esto quiere decir que la varianza no es continua. Esto sucede debido a que en los modelos 2 y 3 se mezclan hombres y mujeres por lo que la región donde están mezclados tiene una varianza mientras que la región donde solo hay hombres tiene otra. Esto no se ve en el modelo 1 porque están separados.

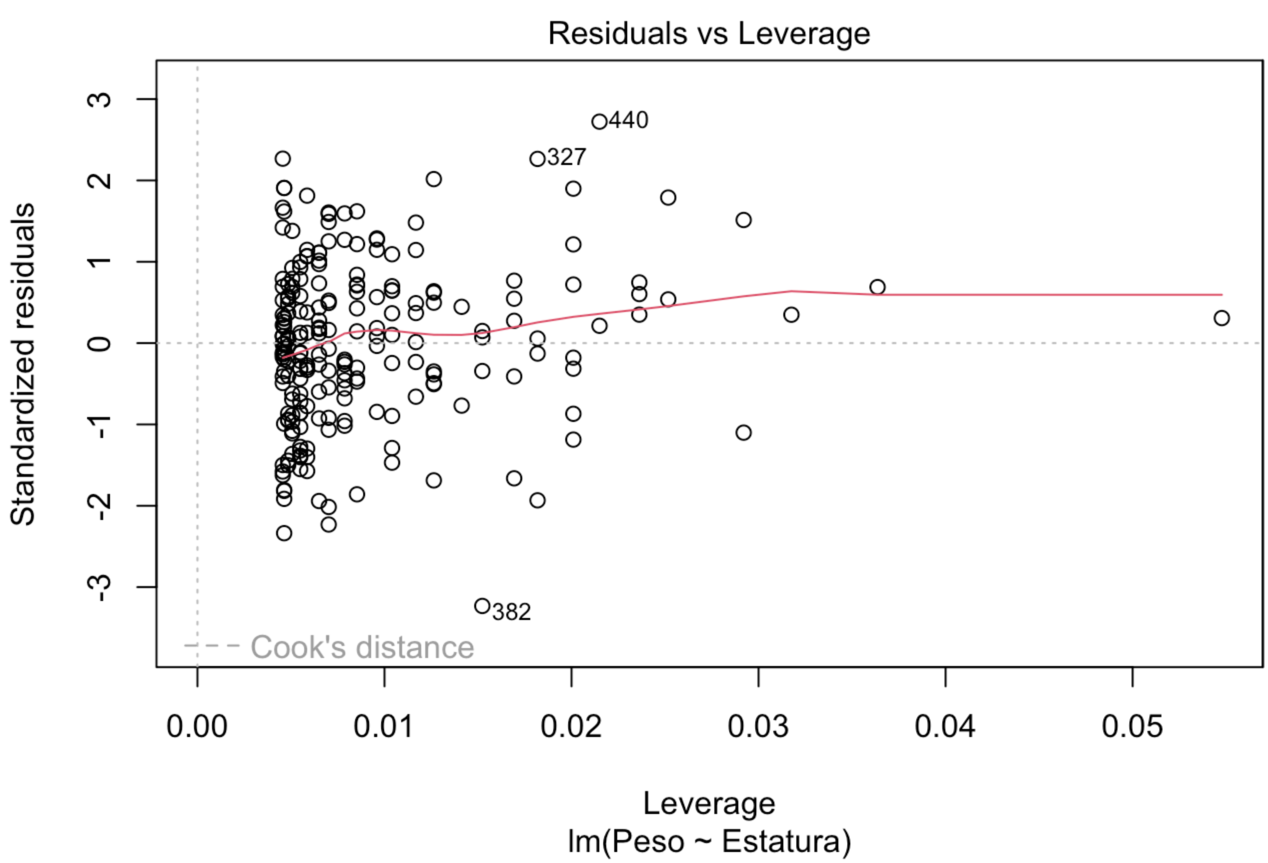
Plot(modelo)

```
plot(Modelo1M)
```

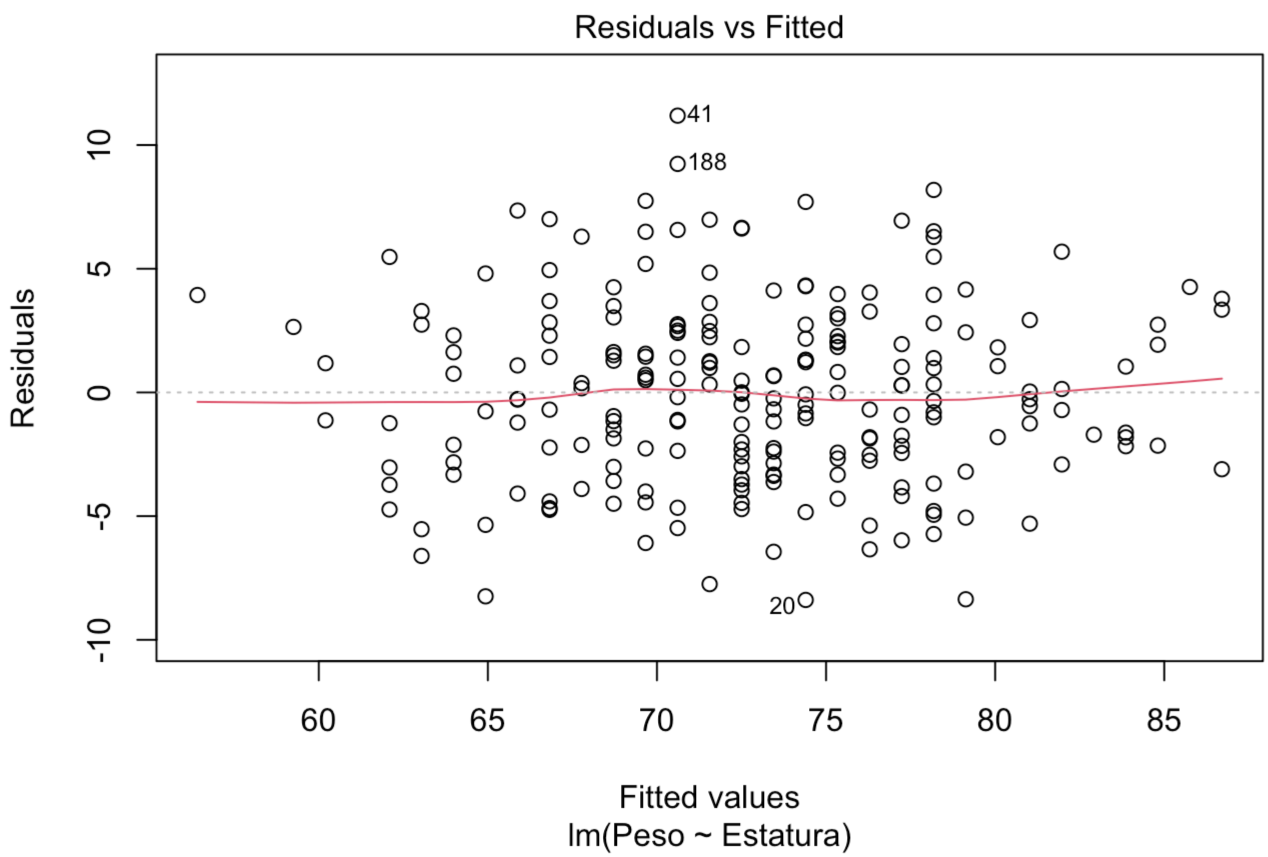

Residuals vs Fitted
Residuals
20
10
0
-10
-20
440
430
382
45
50
55
60
65
Fitted values
lm(Peso ~ Estatura)

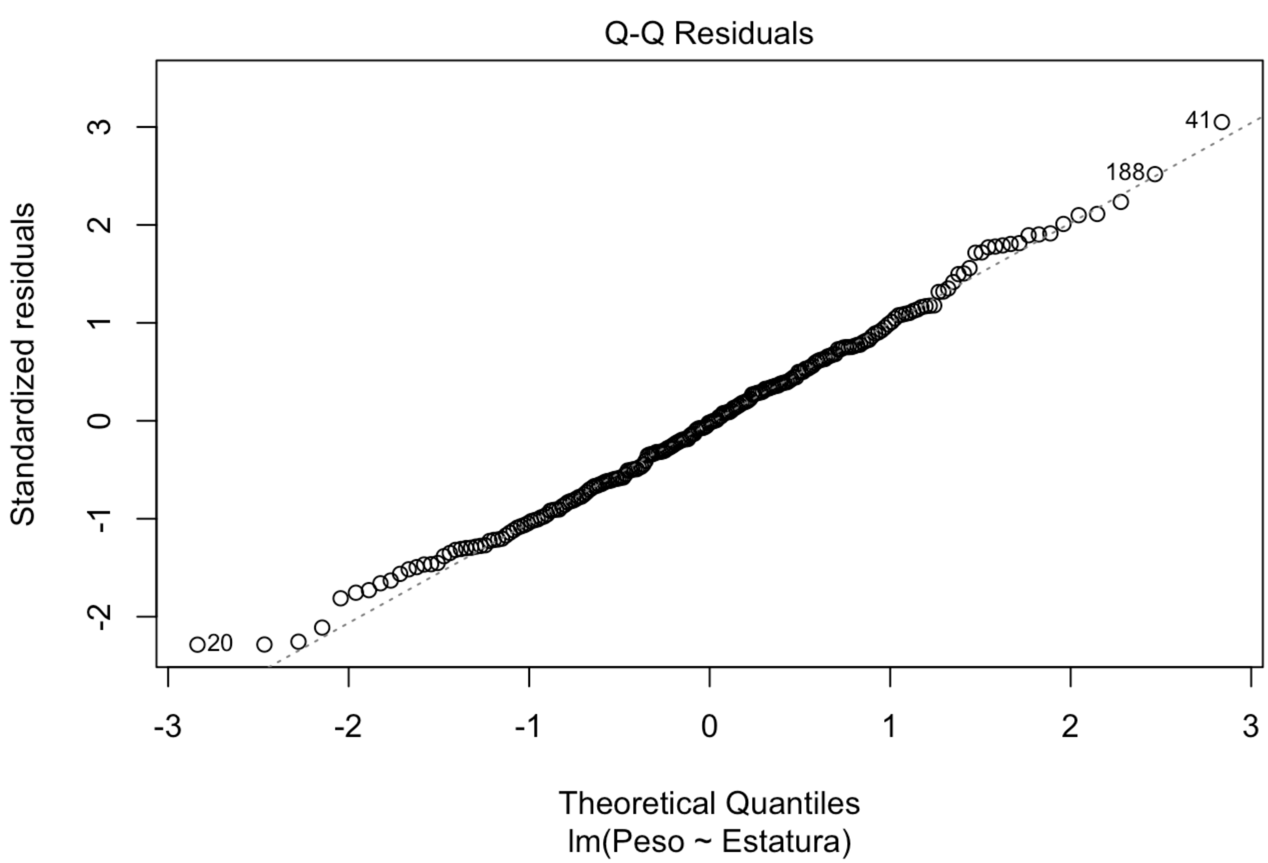

Q-Q Residuals
Standardized residuals
3
2
1
0
-1
-2
-3
440
430
382
Theoretical Quantiles
lm(Peso ~ Estatura)

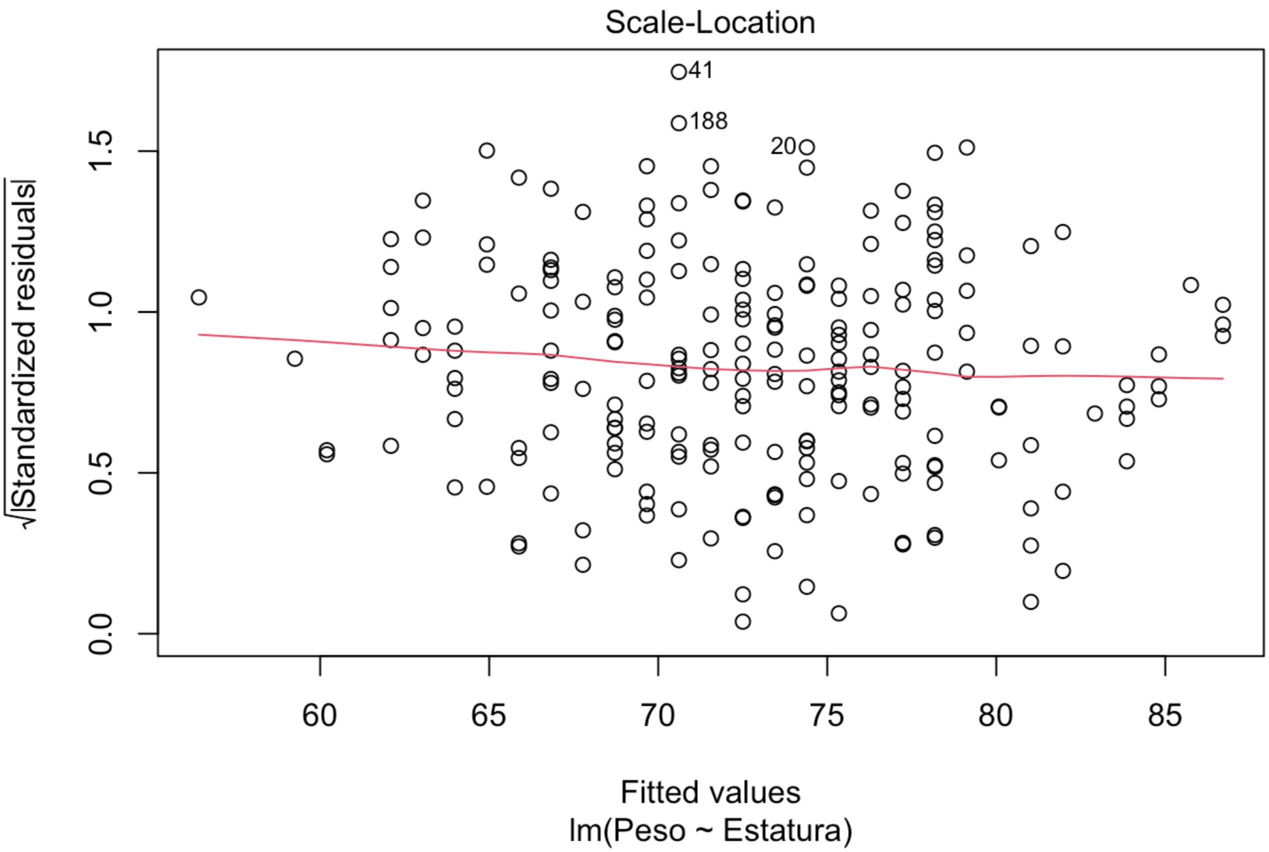

Scale-Location
382
440
430
√|Standardized residuals|
0.0
0.5
1.0
1.5
45
50
55
60
65
Fitted values
lm(Peso ~ Estatura)

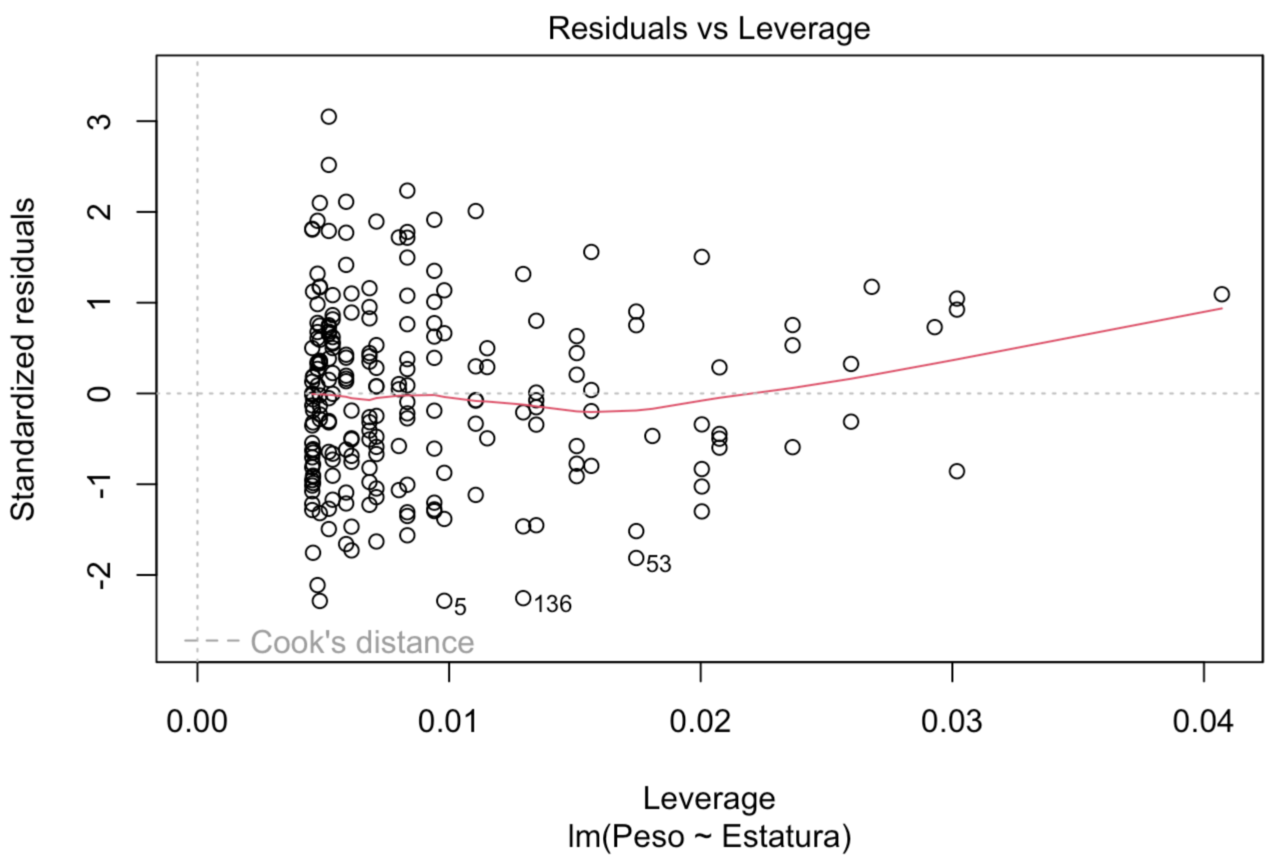

Residuals vs Leverage
440
327
382
Cook's distance
Standardized residuals
-3
-2
-1
0
1
2
3
0.00
0.01
0.02
0.03
0.04
0.05
Leverage
lm(Peso ~ Estatura)

```
plot(Modelo1H)
```

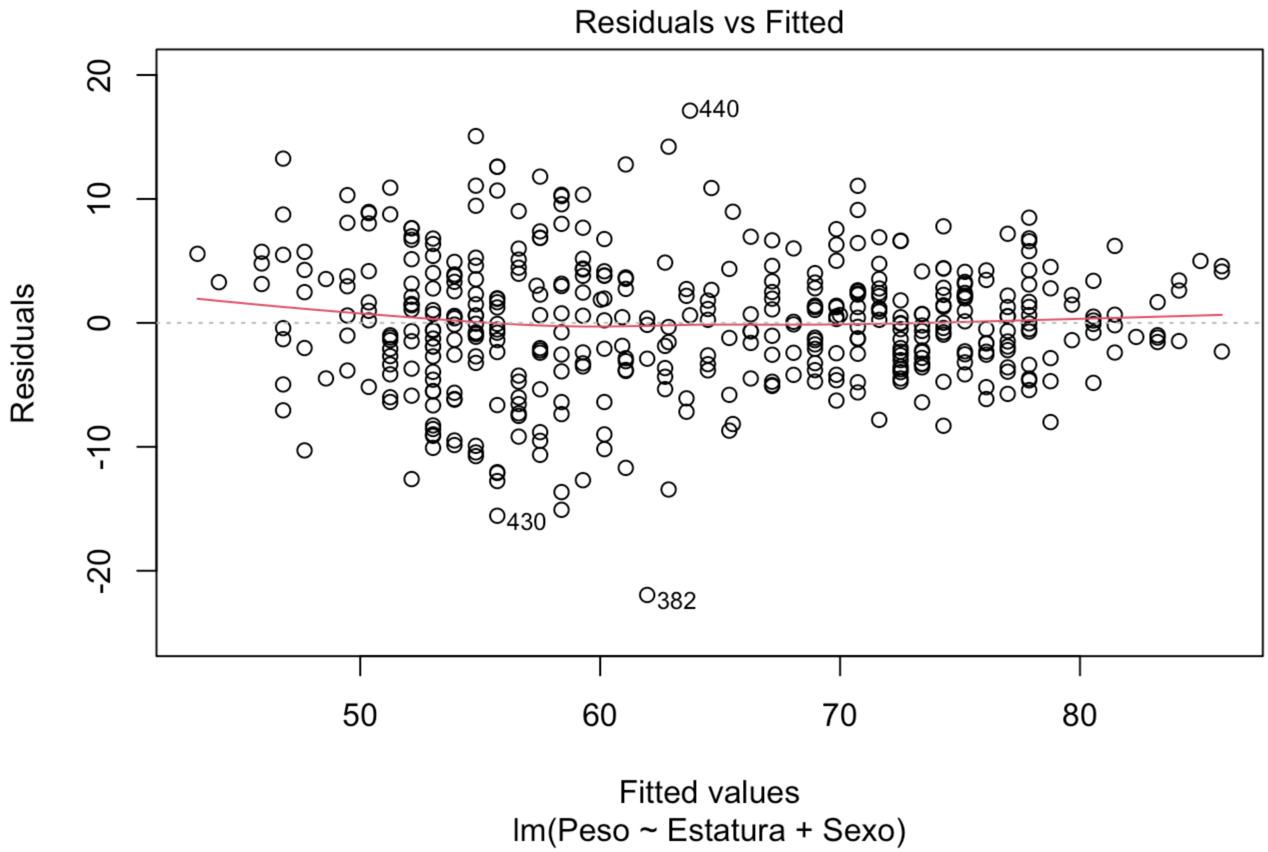

Residuals vs Fitted
Residuals
Fitted values
lm(Peso ~ Estatura)
41
188
20
10
5
0
-5
-10
60
65
70
75
80
85

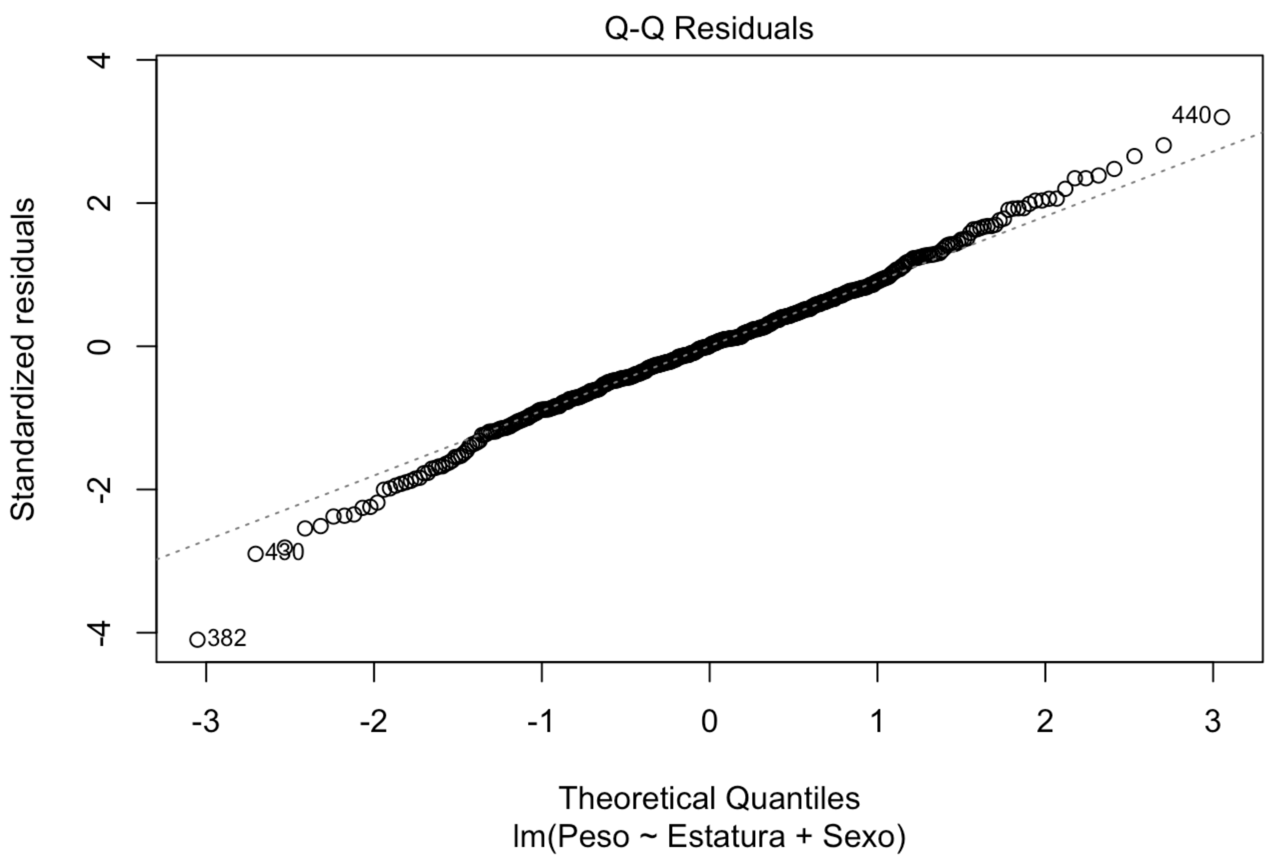

Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Peso ~ Estatura)
41
188
20
3
2
1
0
-1
-2
-3

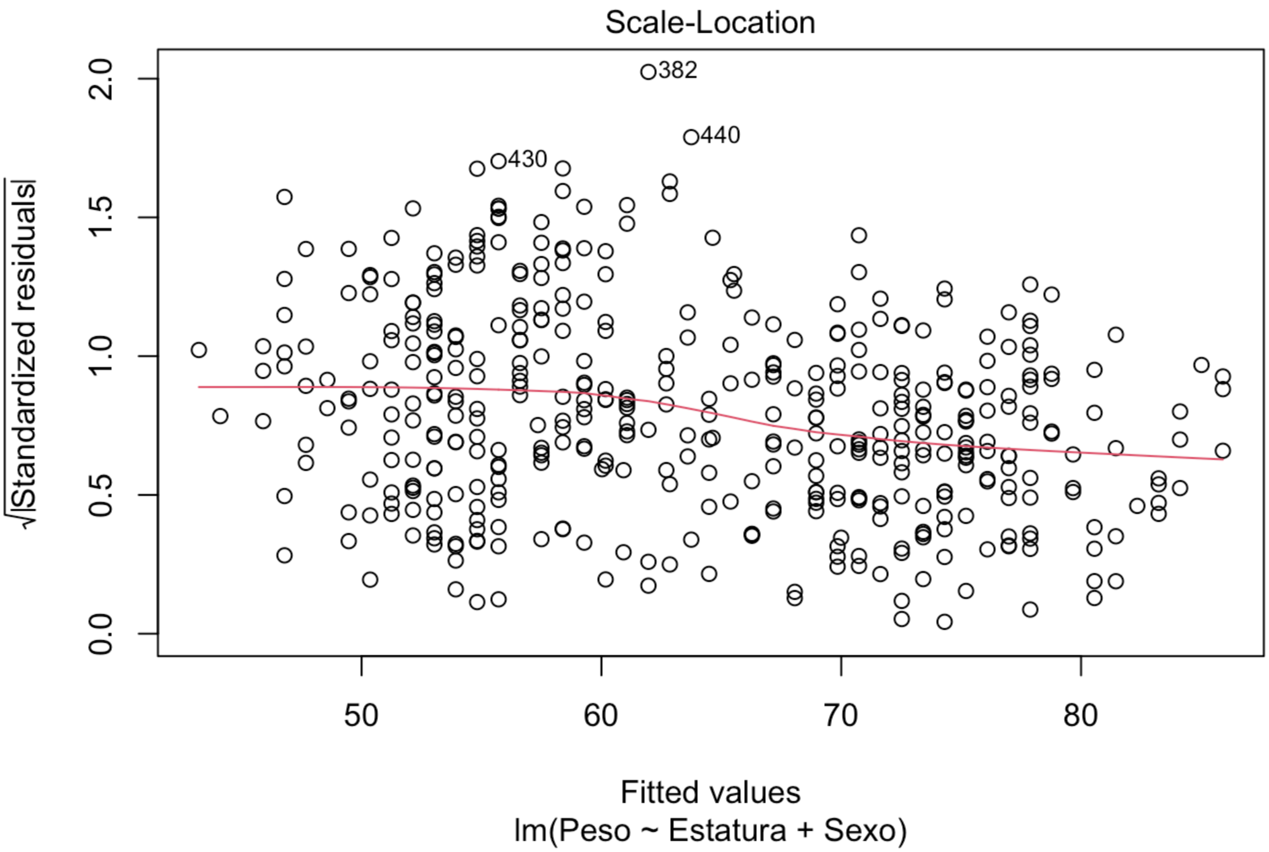

Scale-Location
41
188
20
√|Standardized residuals|
0.0
0.5
1.0
1.5
60
65
70
75
80
85
Fitted values
lm(Peso ~ Estatura)

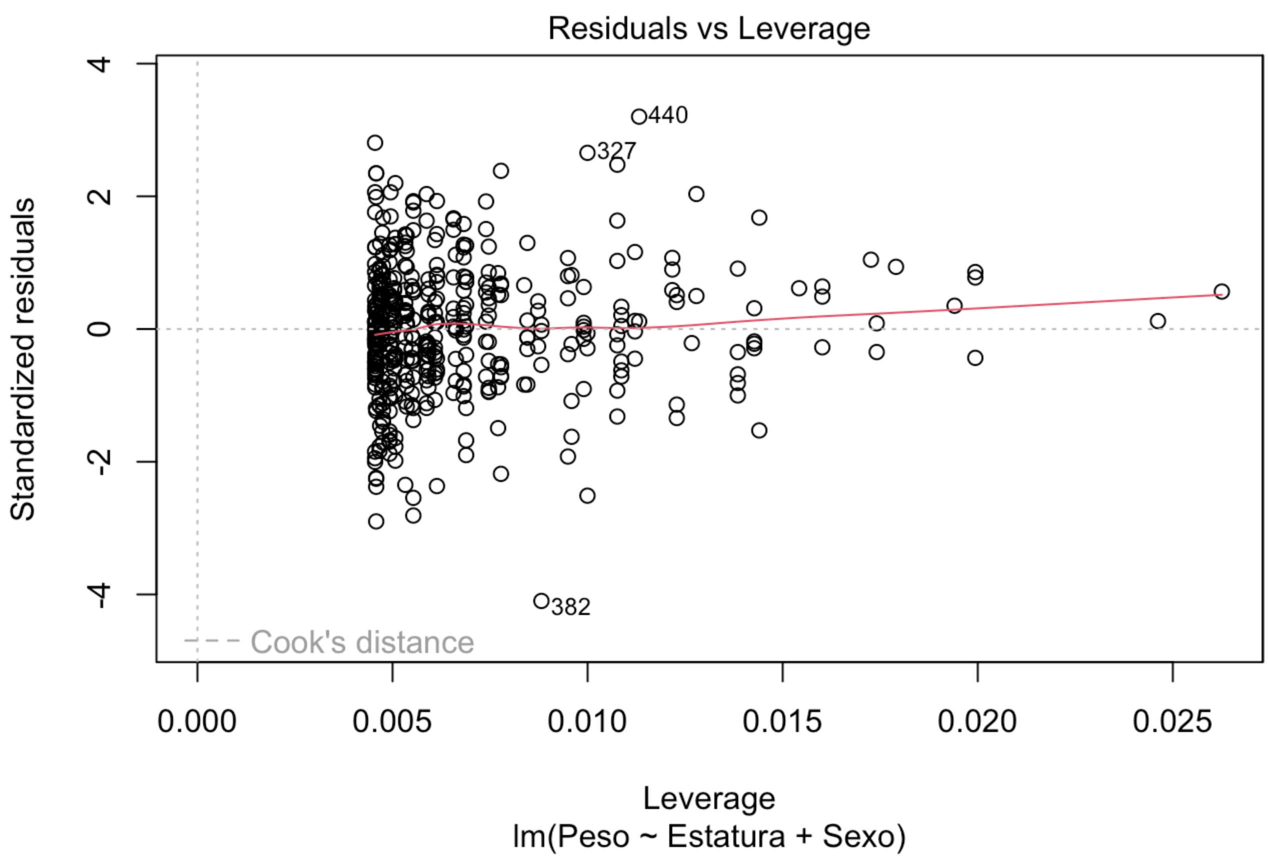

Residuals vs Leverage
Standardized residuals
3
2
1
0
-1
-2
5
136
53
Cook's distance
0.00
0.01
0.02
0.03
0.04
Leverage
lm(Peso ~ Estatura)

```
plot(Modelo2)
```

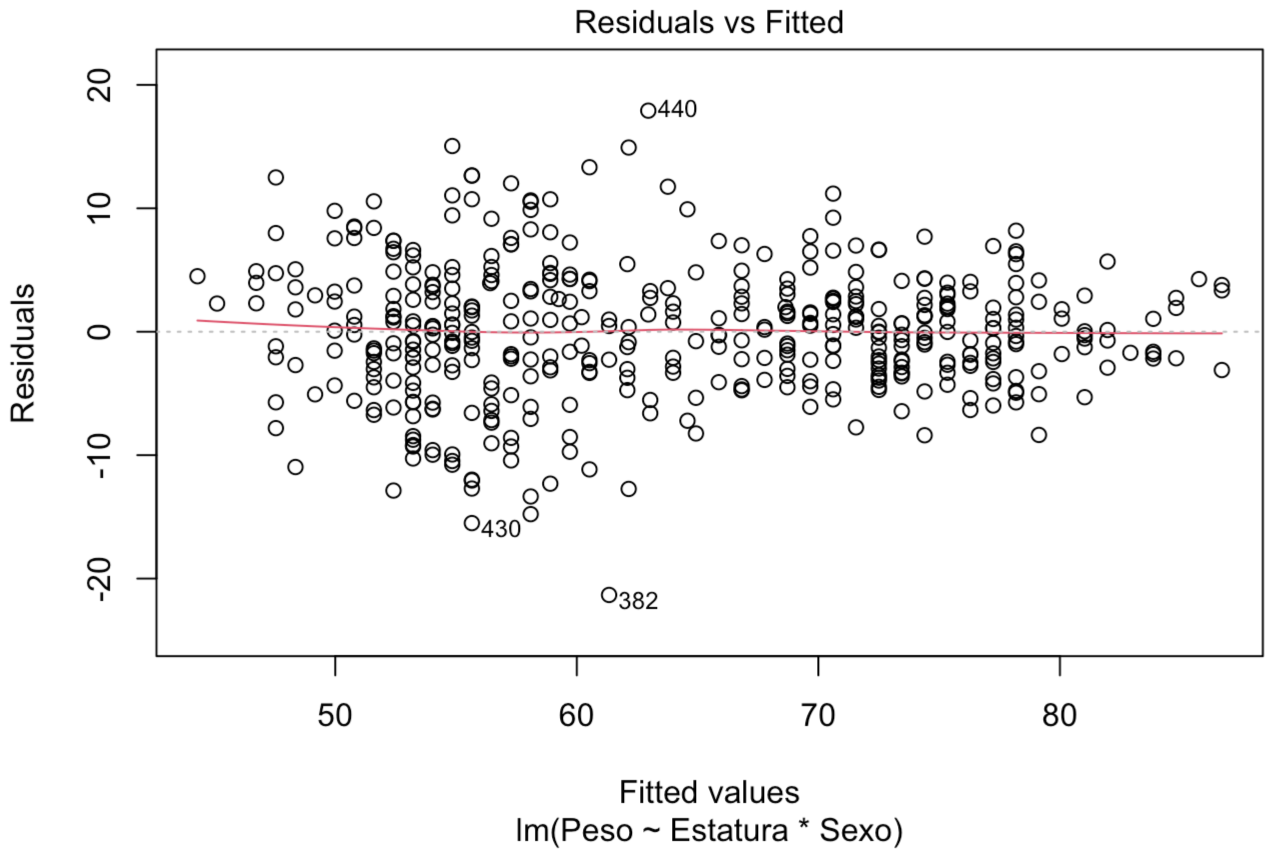

Residuals vs Fitted
Residuals
Fitted values
lm(Peso ~ Estatura + Sexo)
440
430
382
20
10
0
-10
-20
50
60
70
80

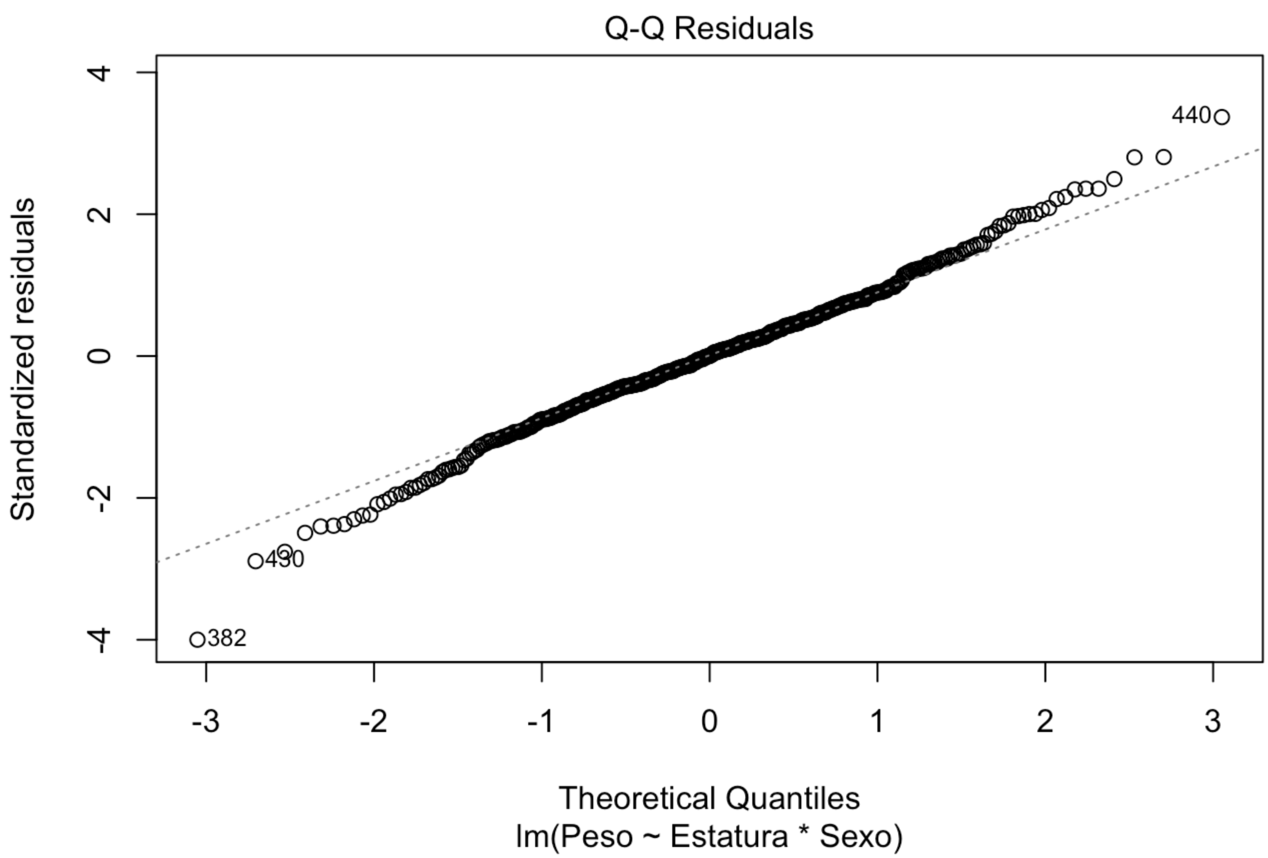

Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Peso ~ Estatura + Sexo)
440
430
382
4
2
0
-2
-4
-3
-2
-1
0
1
2
3

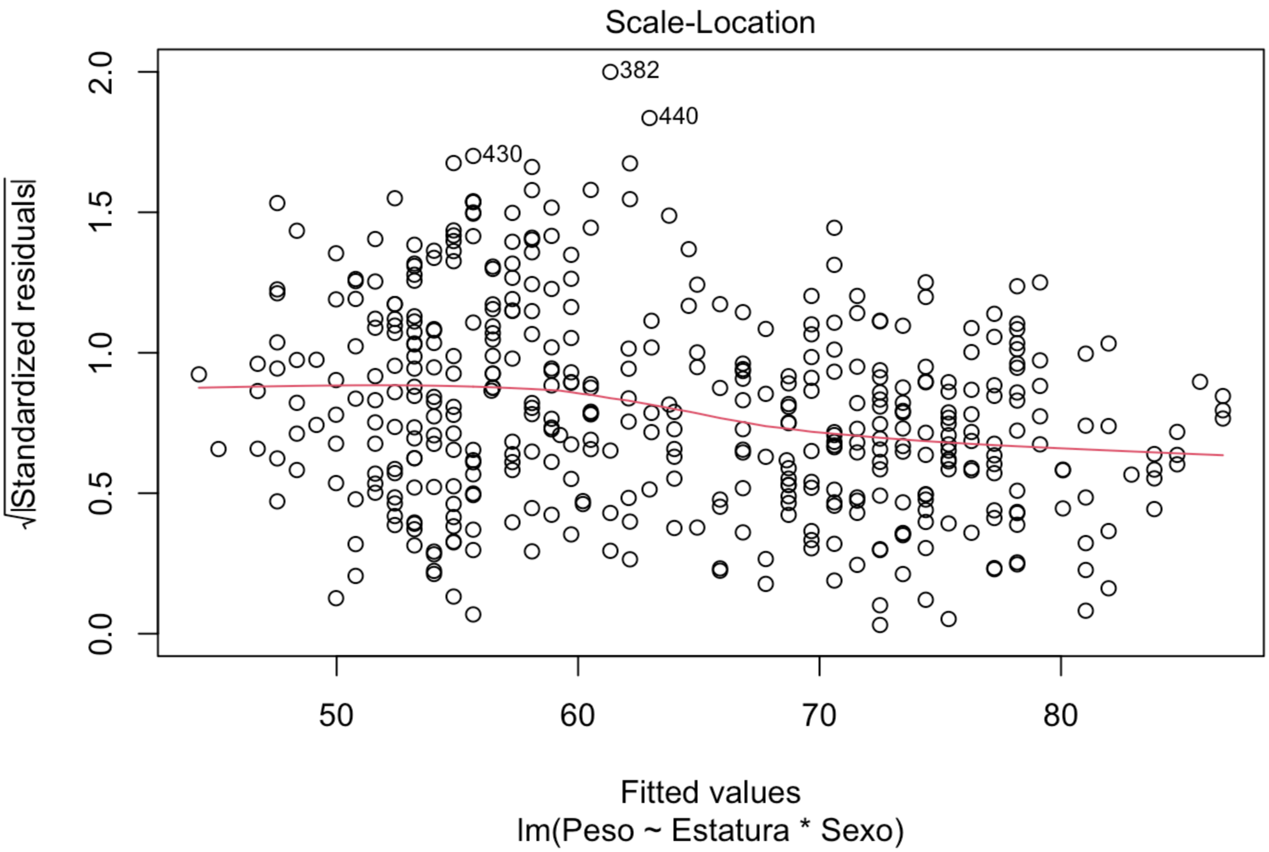

Scale-Location
√|Standardized residuals|
Fitted values
lm(Peso ~ Estatura + Sexo)
382
440
430

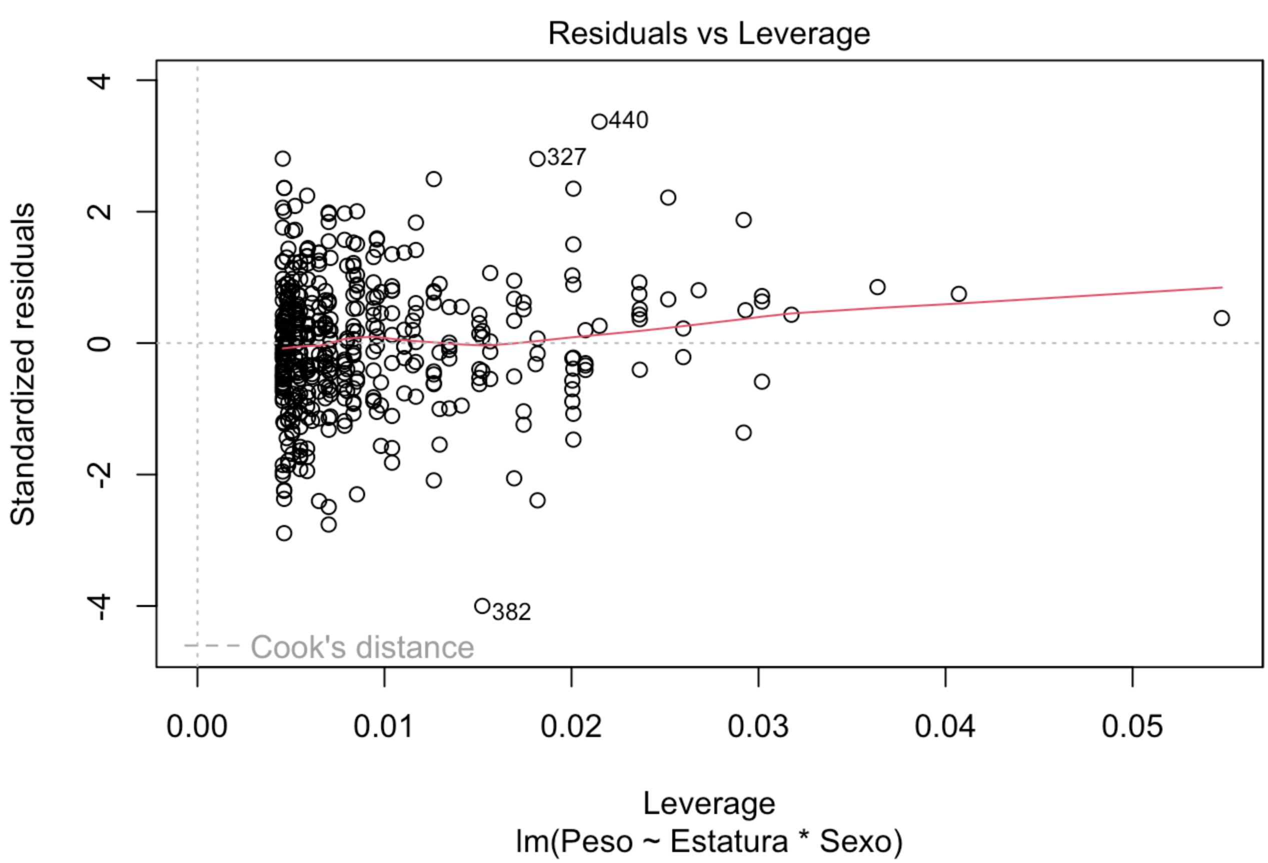

Residuals vs Leverage
Standardized residuals
Leverage
lm(Peso ~ Estatura + Sexo)
440
327
382
Cook's distance

```
plot(Modelo3)
```


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)


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


Residuals vs Leverage
Standardized residuals
Leverage
lm(Peso ~ Estatura * Sexo)
440
327
382
Cook's distance
-4
-2
0
2
4
0.00
0.01
0.02
0.03
0.04
0.05

Estas gráficas y las anteriores proporcionan información similar, analizan los residuos respecto a los valores de predicción y la normalidad de residuos, en este caso también se incluye una gráfica del leverage que permite analizar la importancia de cada dato y su influencia en el modelo, de esta forma se pueden detectar puntos que afecten negativamente al modelo, estos serían puntos con alto leverage y alto residuo.

Estas gráficas no cambian las conclusiones obtenidas anteriormente.

El mejor modelo

El mejor modelo que se obtiene de este análisis es el tercer modelo, este modelo es capaz de realizar una ecuación para explicar la relación entre peso y estatura tanto de hombres como mujeres, donde en esta ecuación se pueden modificar los pesos dependiendo del sexo, diferencia del modelo 1 que se tienen que separar los datos para lograr esto. El modelo 2 no es tan bueno como el 3 porque este solo puede modificar el intercept dependiendo del sexo, mientras que la pendiente es fija, lo que otorga menos flexibilidad para lograr un mejor ajuste.

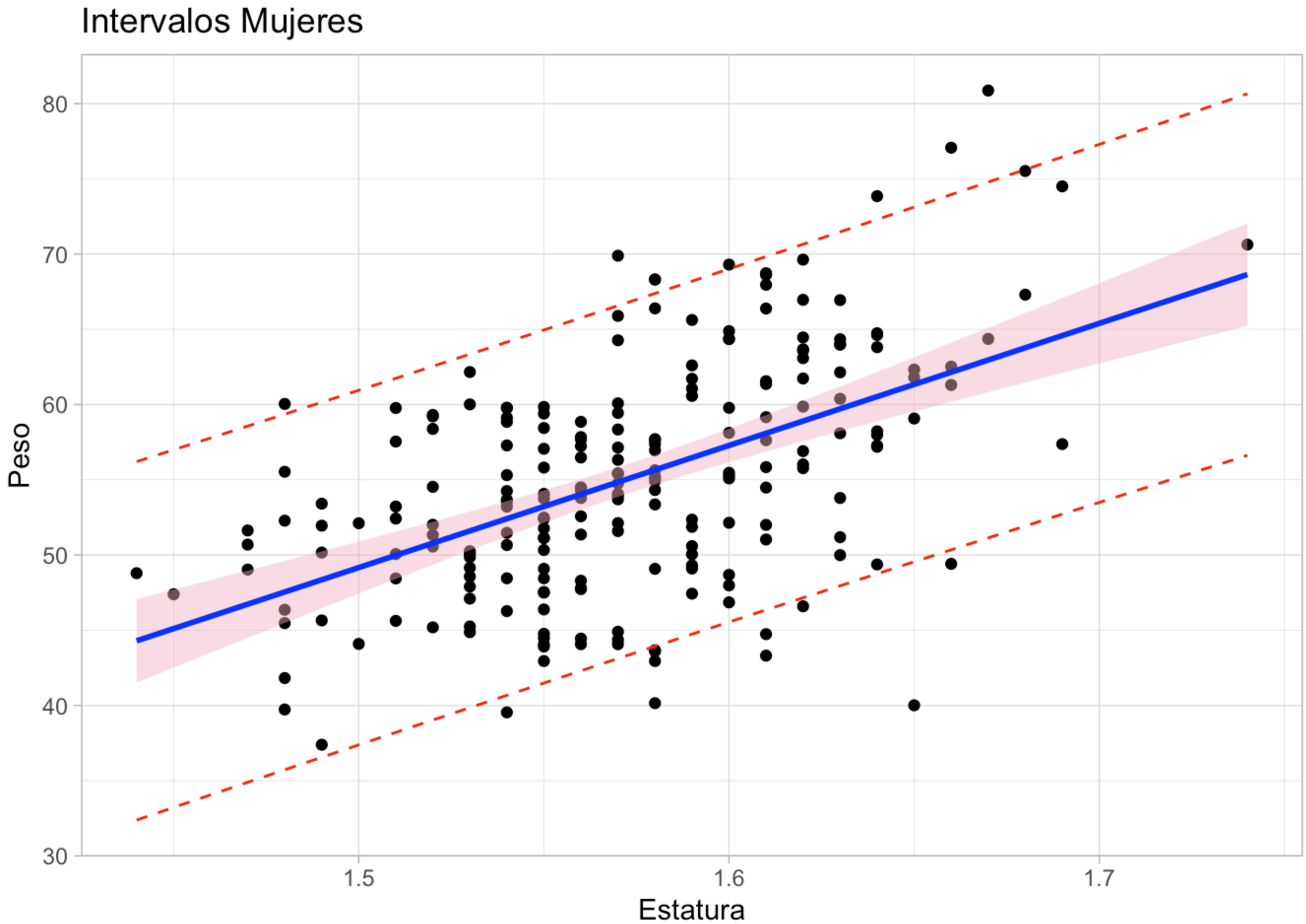
Además el modelo 3 cumple con la mayoría de las condiciones de la regresión lineal, solo falla con la de varianza constante, pero esto no es porque el modelo no sea bueno, sino que es por la diferencia que hay entre hombres y mujeres.

Intervalos de confianza

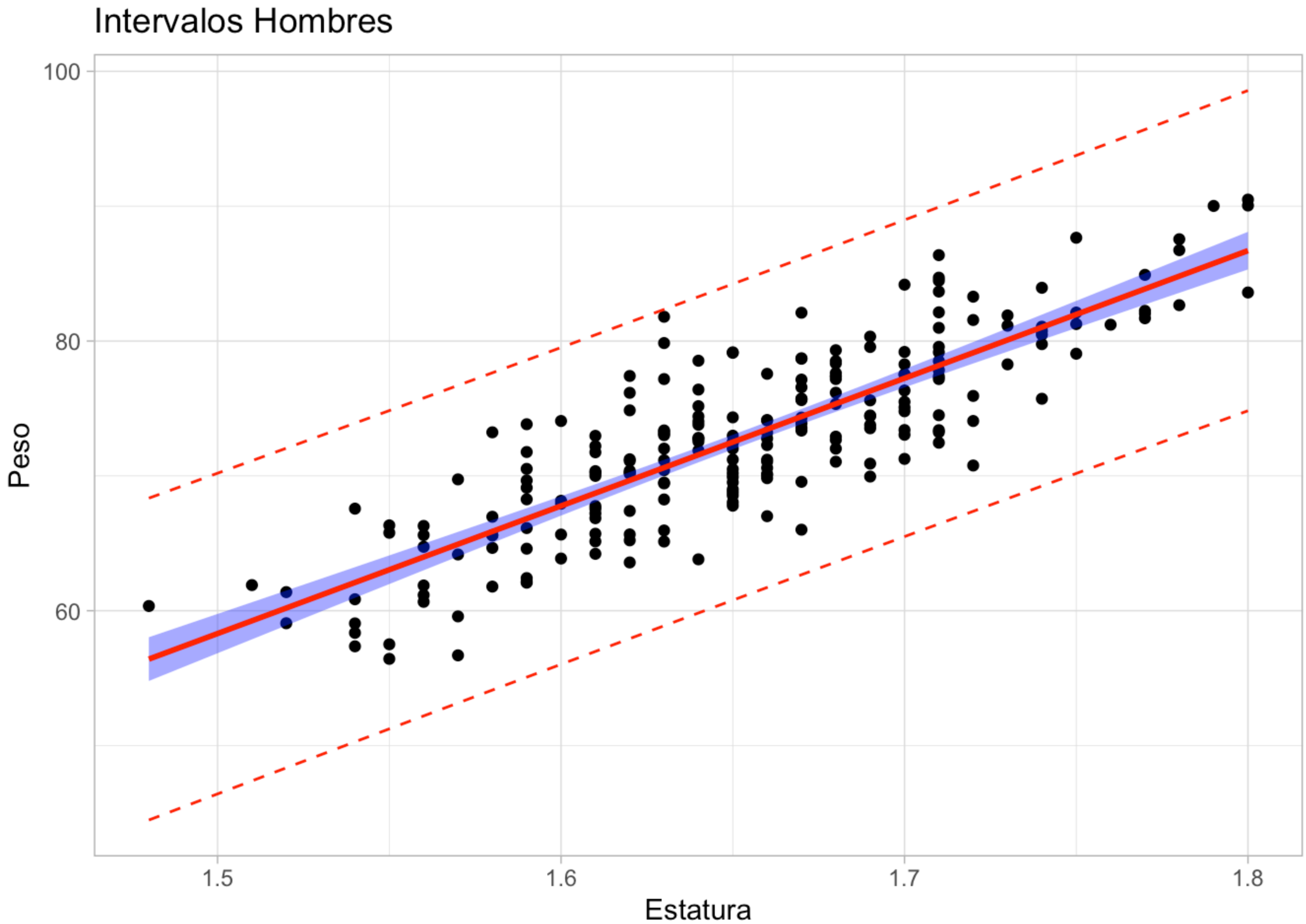
```
Ip=predict(object=Modelo3,interval="prediction",level=0.97)
```

```
## Warning in predict.lm(object = Modelo3, interval = "prediction", level = 0.97): predi
ctions on current data refer to _future_ responses
```

```
datos1=cbind(M,Ip)
MM =subset(datos1, M$Sexo=='M')
MH =subset(datos1, M$Sexo=='H')

library(ggplot2)
ggplot(MM,aes(x=Estatura,y=Peso),)+
  ggtitle("Intervalos Mujeres")+
  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()
```




```
ggplot(MH,aes(x=Estatura,y=Peso),)+
  ggtitle("Intervalos Hombres")+
  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="red", fill="blue")+
  theme_light()
```



Se dividieron en dos gráficas para poder visualizar el comportamiento de hombres y mujeres, se observa que el intervalo de confianza para mujeres es un poco más ancho debido a que hay más variación en los datos. Esto quiere decir que las predicciones hechas por el modelo de hombres son más acertadas.