12. Regresión Lineal - Análisis de los errores

Ozner Leyva

2024-09-04

Analiza si el (los) modelo(s) obtenidos anteriormente son apropiados para el conjunto de datos. Realiza el análisis de los residuos:

```
M = read.csv("C:/Users/ozner/Desktop/Computer Science/R/Estatura-peso_HyM.csv")
MM = subset(M,M$Sexo=="M")
MH = subset(M,M$Sexo=="H")
Modelo1H = lm(Estatura~Peso, MH)
Modelo1M = lm(Estatura~Peso, MM)
Modelo2 = lm(Estatura~Peso, M)
Modelo3 = lm(Peso~Estatura*Sexo, M)
```

Normalidad de los residuos

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

```
##
##  Anderson-Darling normality test
##
## data:  Modelo1H$residuals
## A = 0.38581, p-value = 0.3884
```

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

```
##
##  Anderson-Darling normality test
##
## data:  Modelo1M$residuals
## A = 0.19471, p-value = 0.8909
```

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

```
##
##  Anderson-Darling normality test
##
## data:  Modelo2$residuals
## A = 0.25276, p-value = 0.7341
```

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

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

Hipótesis nula (H_0): Los datos siguen una distribución normal. Hipótesis alternativa (H_1): Los datos no siguen una distribución normal.

Si el valor p es bajo (< 0.05), se rechaza la hipótesis nula, lo que sugiere que los datos no son normales.

Entonces solo hay normalidad en los modelos: -Modelo1H -Modelo1M -Modelo2

Verificación de media cero

```
t.test(Modelo1H$residuals)

##
##  One Sample t-test
##
## data:  Modelo1H$residuals
## t = -4.3085e-16, df = 219, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.004362545  0.004362545
## sample estimates:
##     mean of x
## -9.536975e-19

t.test(Modelo1M$residuals)

##
##  One Sample t-test
##
## data:  Modelo1M$residuals
## t = 2.1668e-16, df = 219, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.00569816  0.00569816
## sample estimates:
##   mean of x
## 6.26481e-19

t.test(Modelo2$residuals)

##
##  One Sample t-test
##
## data:  Modelo2$residuals
## t = 5.1736e-18, df = 439, p-value = 1
## alternative hypothesis: true mean is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.003867162  0.003867162
## sample estimates:
##   mean of x
## 1.01798e-20
```

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

```
##
##  One Sample t-test
##
## data:  Modelo3$residuals
## t = 8.3074e-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
## 2.120936e-16
```

Hipótesis nula (H_0): La media de los residuos es igual a 0. Hipótesis alternativa (H_1): La media de los residuos es diferente de 0.

Si el valor p resultante es menor que el nivel de significancia (0.05), se rechaza la hipótesis nula, lo que sugiere que la media de los residuos es significativamente diferente de 0.

Entonces, todos los modelos tienen media de 0.

Homocedasticidad e 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(Modelo1H)
```

```
##
##  Durbin-Watson test
##
## data:  Modelo1H
```

```
## DW = 1.9884, p-value = 0.4659
## 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 = 0.0010484, df = 1, p-value = 0.9742
```

```
dwtest(Modelo1M)
```

```
##
##  Durbin-Watson test
##
## data:  Modelo1M
## DW = 2.0825, p-value = 0.7297
## 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 = 0.39351, df = 1, p-value = 0.5305
```

```
dwtest(Modelo2)
```

```
##
##  Durbin-Watson test
##
## data:  Modelo2
## DW = 1.9578, p-value = 0.3184
## 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 = 0.16657, df = 1, p-value = 0.6832
```

```
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
```

De acuerdo a la autocorrelación, el Modelo 3 está descartado.

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

```
bptest(Modelo1H)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Modelo1H
## BP = 0.45776, df = 1, p-value = 0.4987
```

```
gqtest(Modelo1H)
```

```
##
##  Goldfeld-Quandt test
##
## data:  Modelo1H
## GQ = 0.91478, df1 = 108, df2 = 108, p-value = 0.6778
## alternative hypothesis: variance increases from segment 1 to 2
```

```
bptest(Modelo1M)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Modelo1M
## BP = 1.9859, df = 1, p-value = 0.1588
```

```
gqtest(Modelo1M)
```

```
##
##  Goldfeld-Quandt test
##
## data:  Modelo1M
## GQ = 0.94892, df1 = 108, df2 = 108, p-value = 0.6071
## alternative hypothesis: variance increases from segment 1 to 2
```

```
bptest(Modelo2)
```

```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  Modelo2
## BP = 9.9492, df = 1, p-value = 0.001609
```

```
gqtest(Modelo2)
```

```
##
##  Goldfeld-Quandt test
##
## data:  Modelo2
## GQ = 1.706, df1 = 218, df2 = 218, p-value = 4.502e-05
## alternative hypothesis: variance increases from segment 1 to 2
```

```
bptest(Modelo3)
```

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

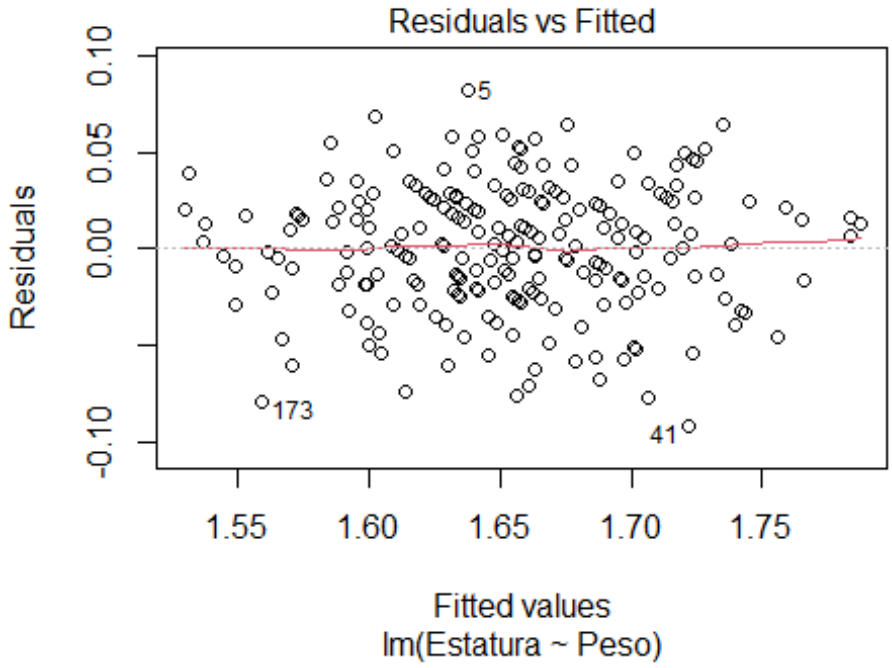
```
gqtest(Modelo3)
```

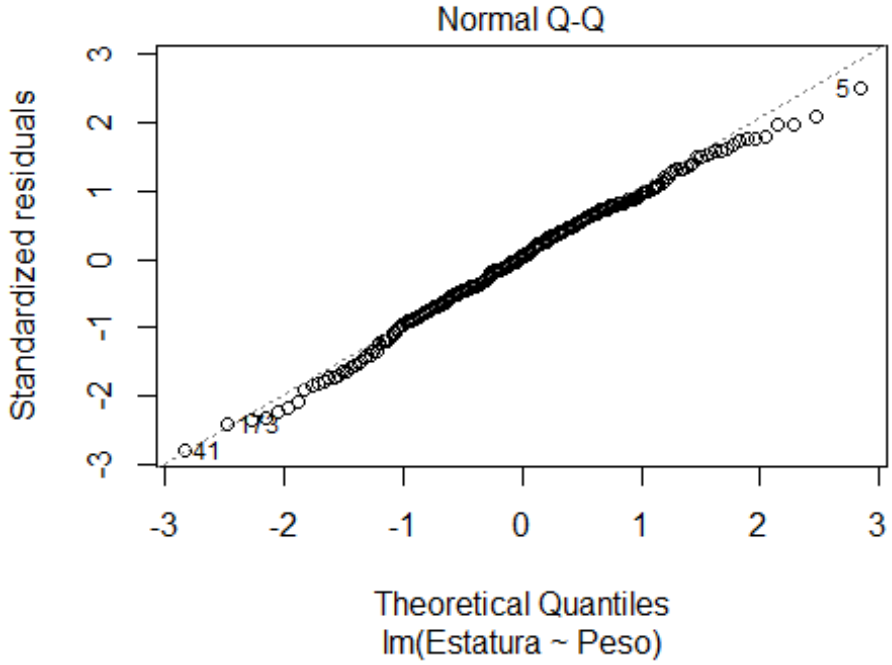
```
##
##  Goldfeld-Quandt test
##
## data:  Modelo3
## GQ = 3.2684, df1 = 216, df2 = 216, p-value < 2.2e-16
## alternative hypothesis: variance increases from segment 1 to 2
```

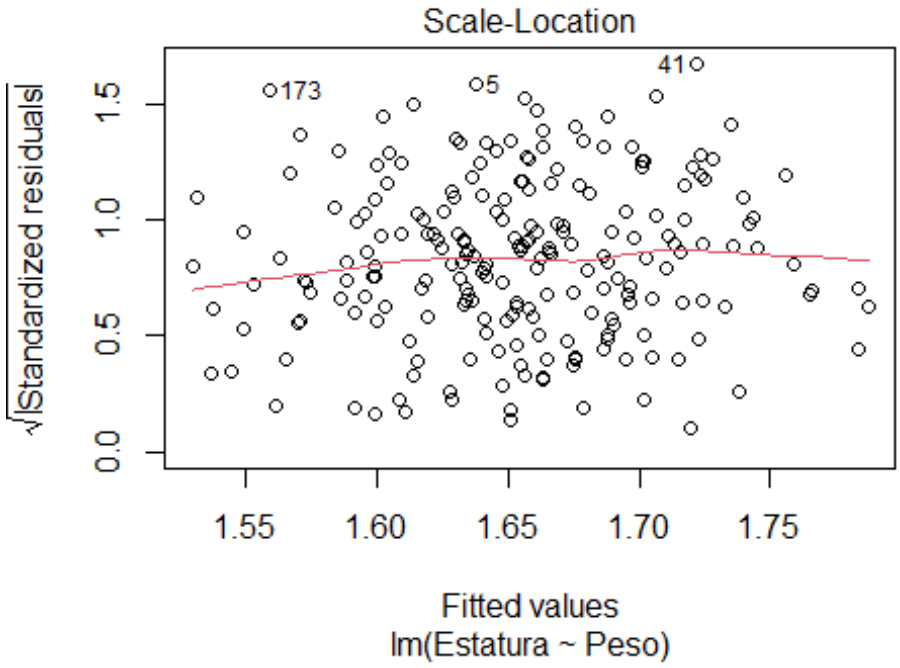
De acuerdo a los tests, los modelos 2 y 3 tienen heterocedasticidad

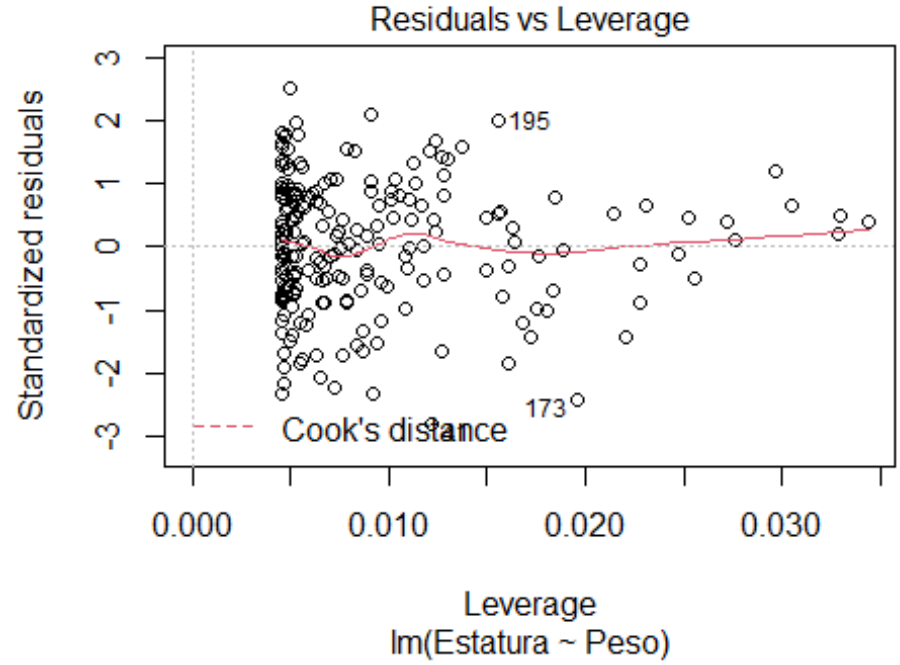
Utiliza el comando: plot(modelo). Observa las gráficas obtenidas y contesta:

```
plot(Modelo1H)
```

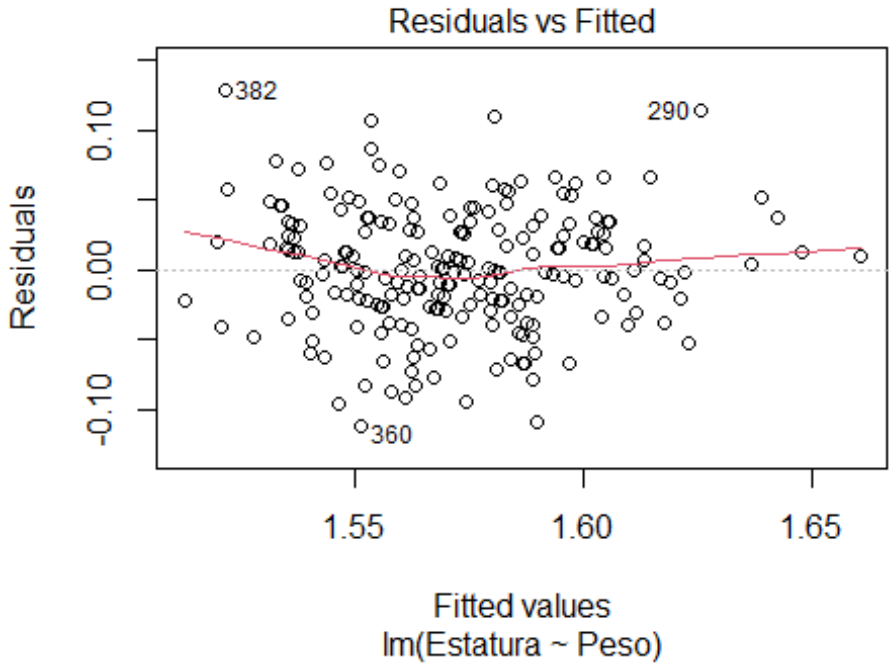

Residuals vs Fitted
Residuals
0.10
0.05
0.00
-0.10
5
173
41
1.55
1.60
1.65
1.70
1.75
Fitted values
lm(Estatura ~ Peso)

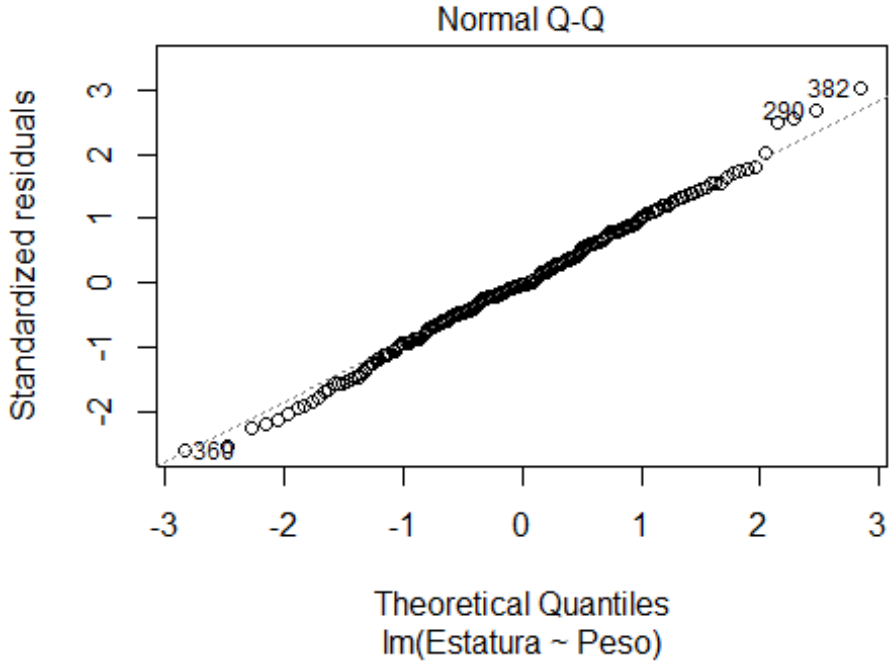

Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
5
173
41
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Estatura ~ Peso)

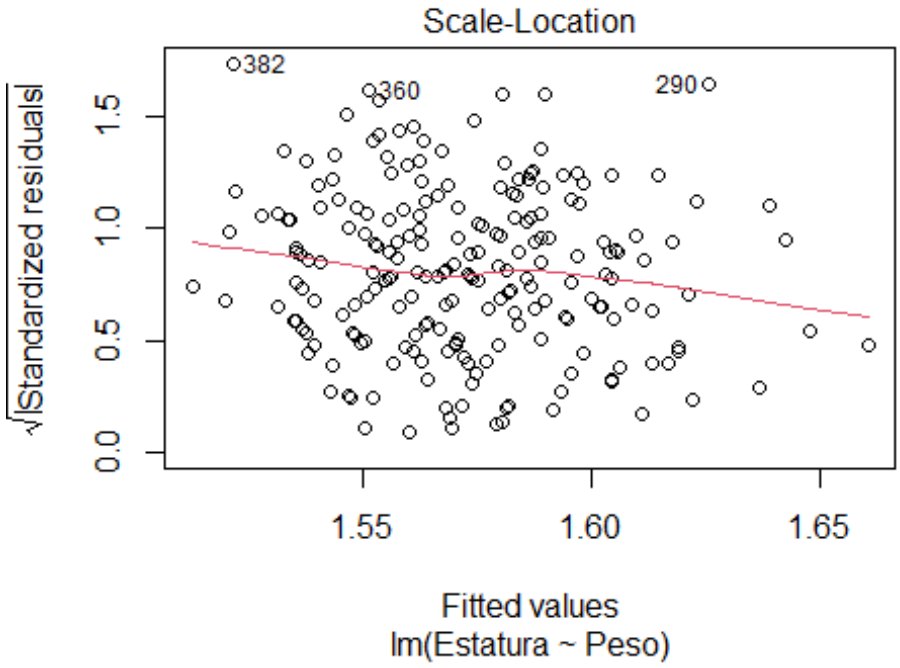


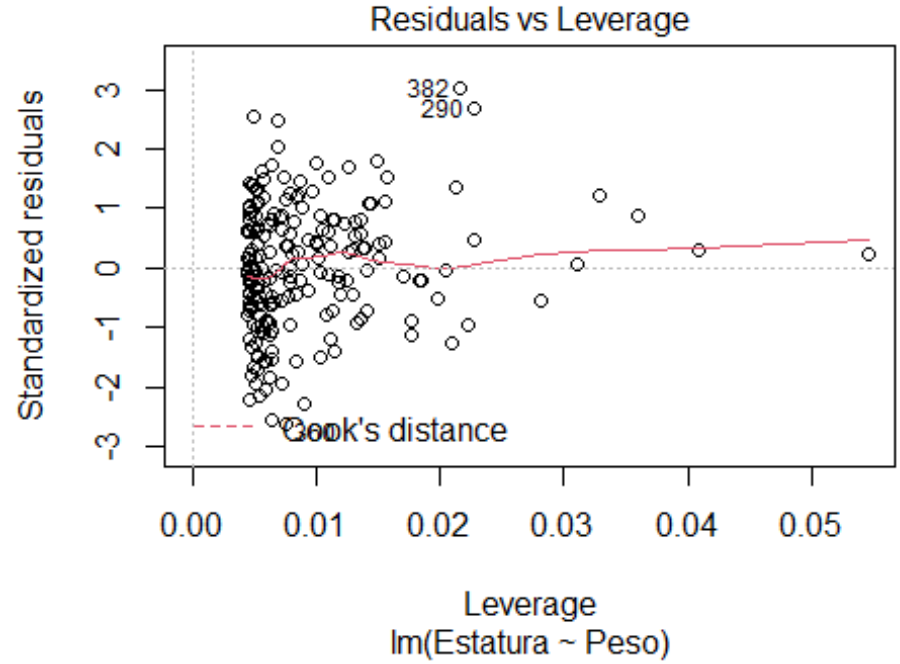



```
plot(Modelo1M)
```

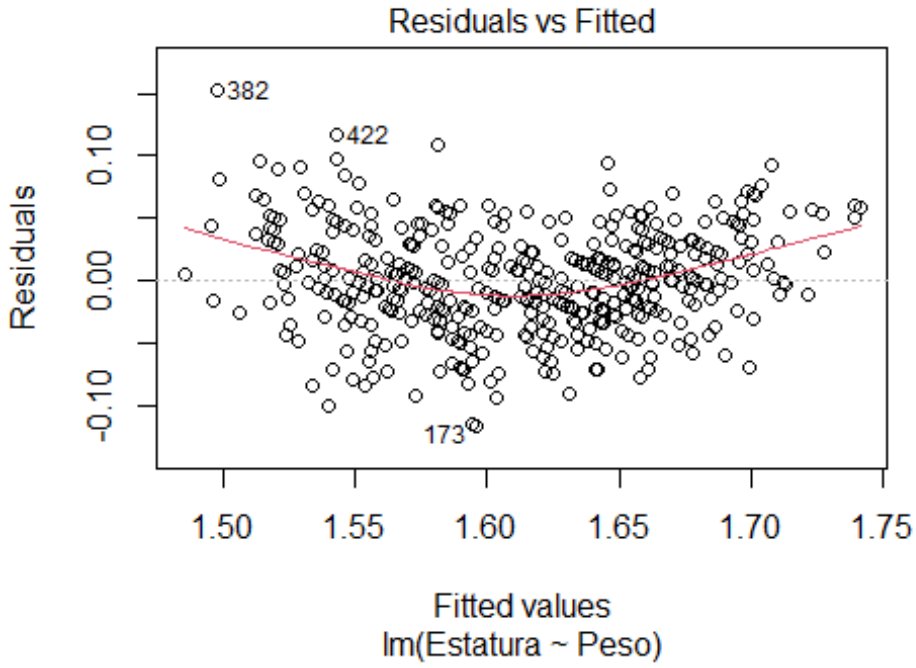

Residuals vs Fitted
Residuals
0.10
0.00
-0.10
382
290
360
1.55
1.60
1.65
Fitted values
lm(Estatura ~ Peso)

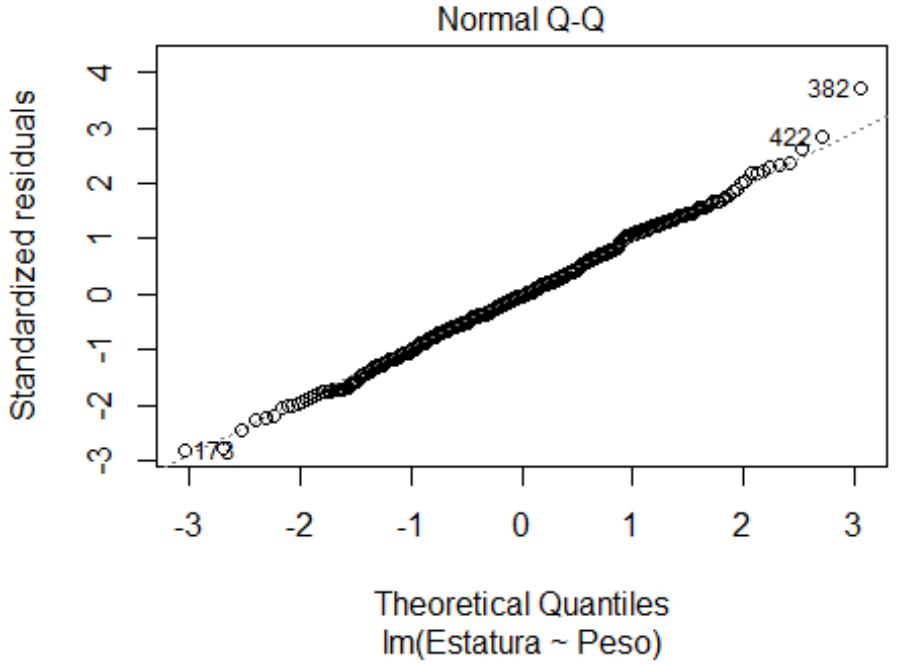

Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
382
290
360
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Estatura ~ Peso)

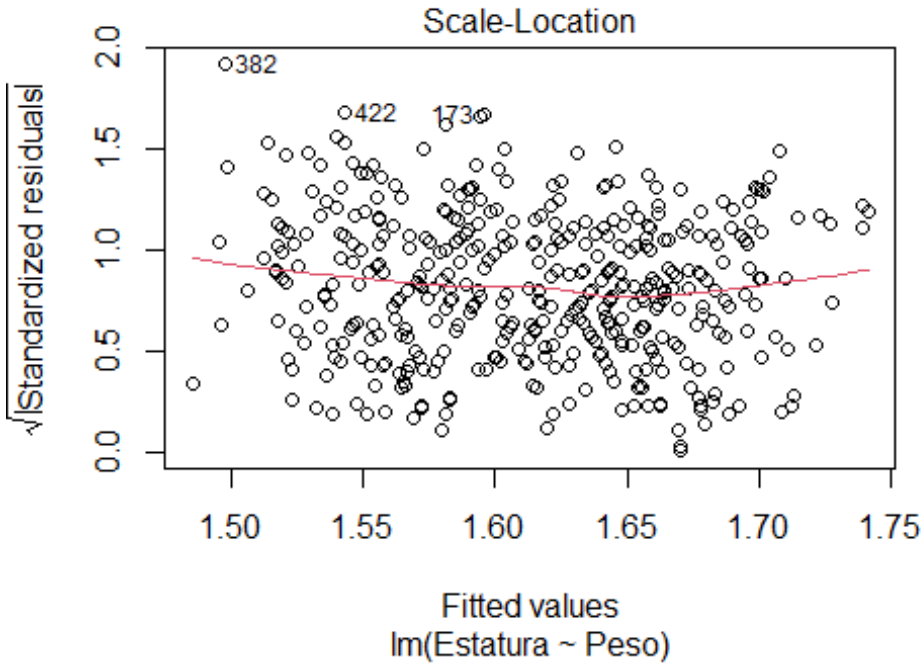


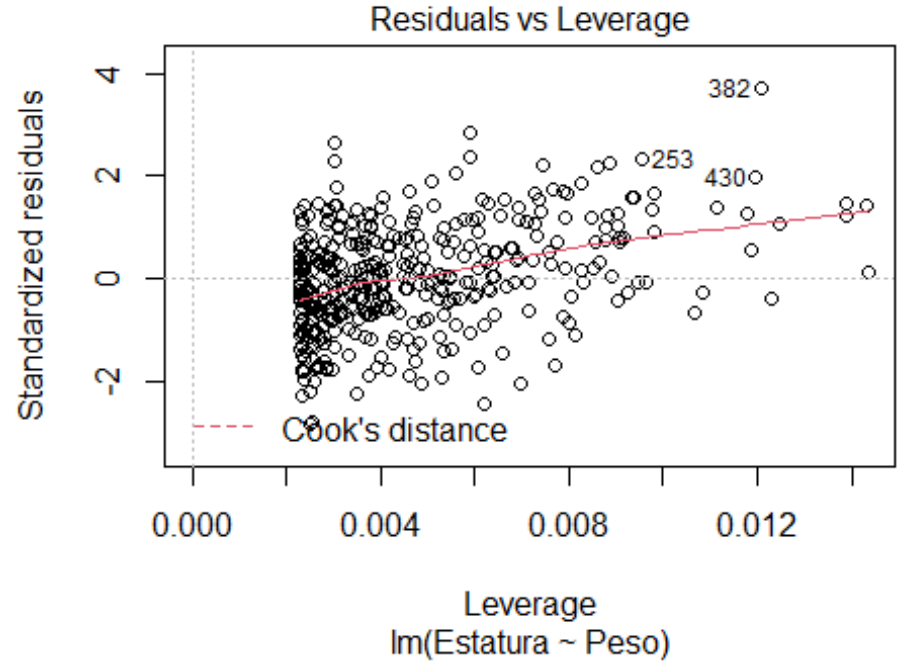



```
plot(Modelo2)
```

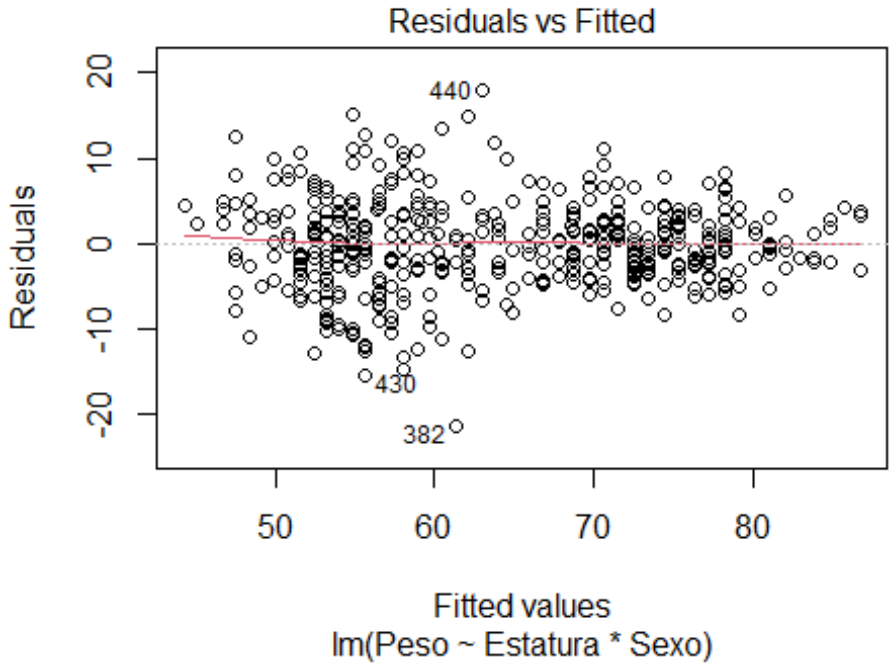

Residuals vs Fitted
382
422
173
Residuals
0.10
0.00
-0.10
1.50
1.55
1.60
1.65
1.70
1.75
Fitted values
lm(Estatura ~ Peso)

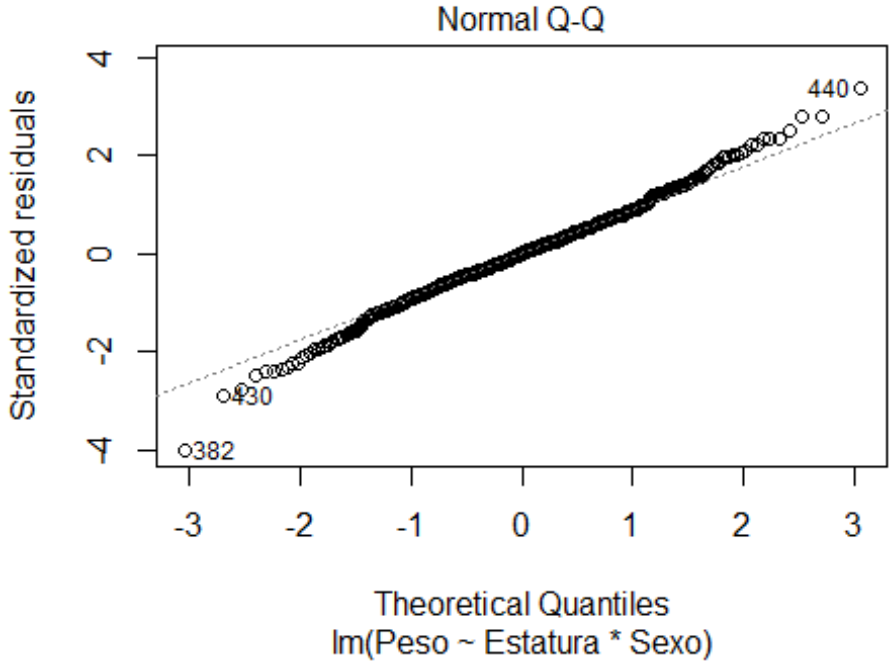

Normal Q-Q
382
422
173
Standardized residuals
4
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Estatura ~ Peso)

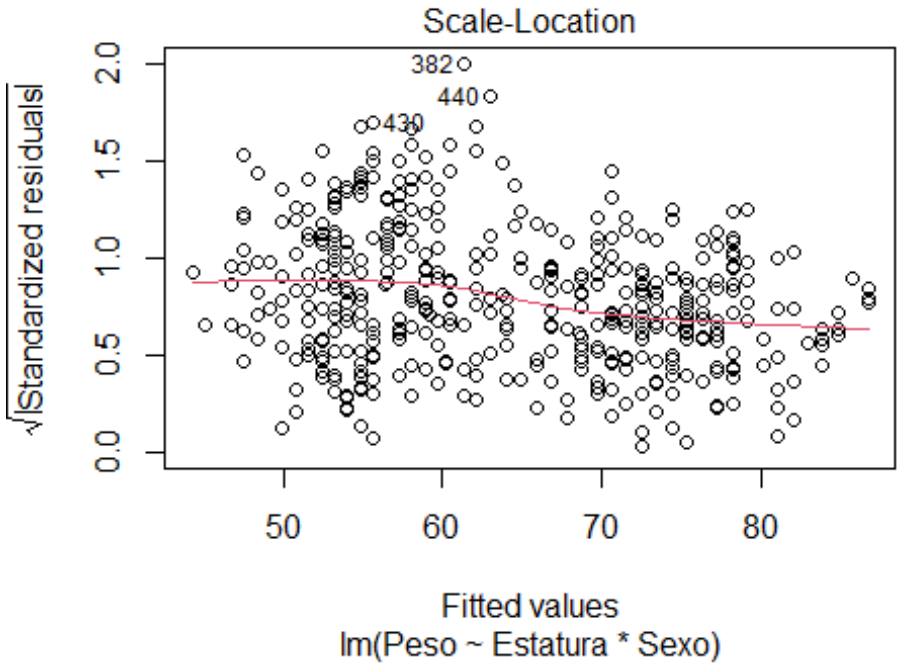


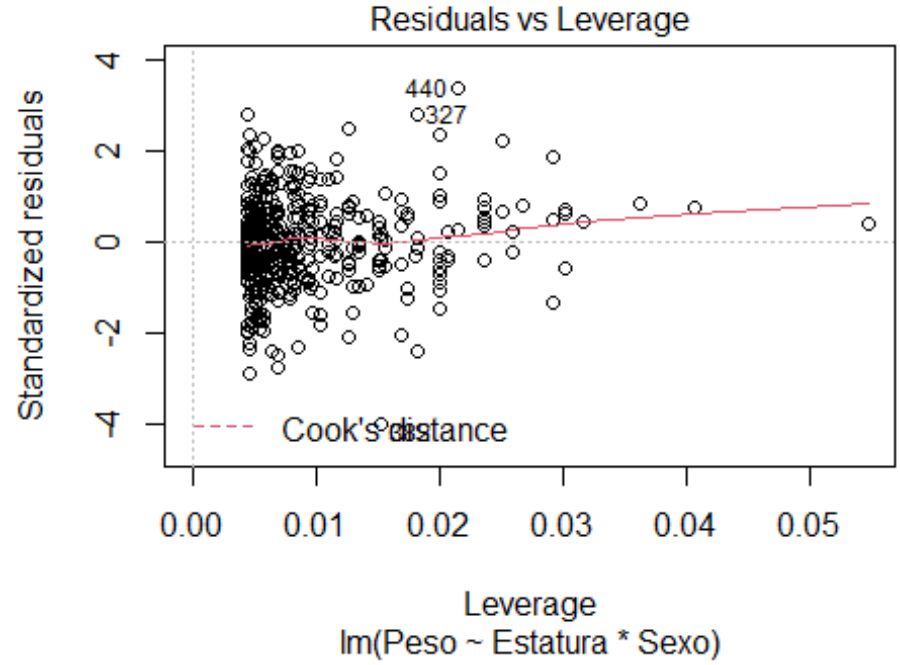



```
plot(Modelo3)
```


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
4
2
0
-2
-4
440
430
382
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Peso ~ Estatura * Sexo)


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)


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

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

Las escalas, estos gráficos exageran las escalas para poder visualizar mejor los errores.

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

No cambian mis conclusiones, pero puedo ver que a pesar de que se rechazan algunas hipotesis, en las graficas no hay una diferencia exagerada.

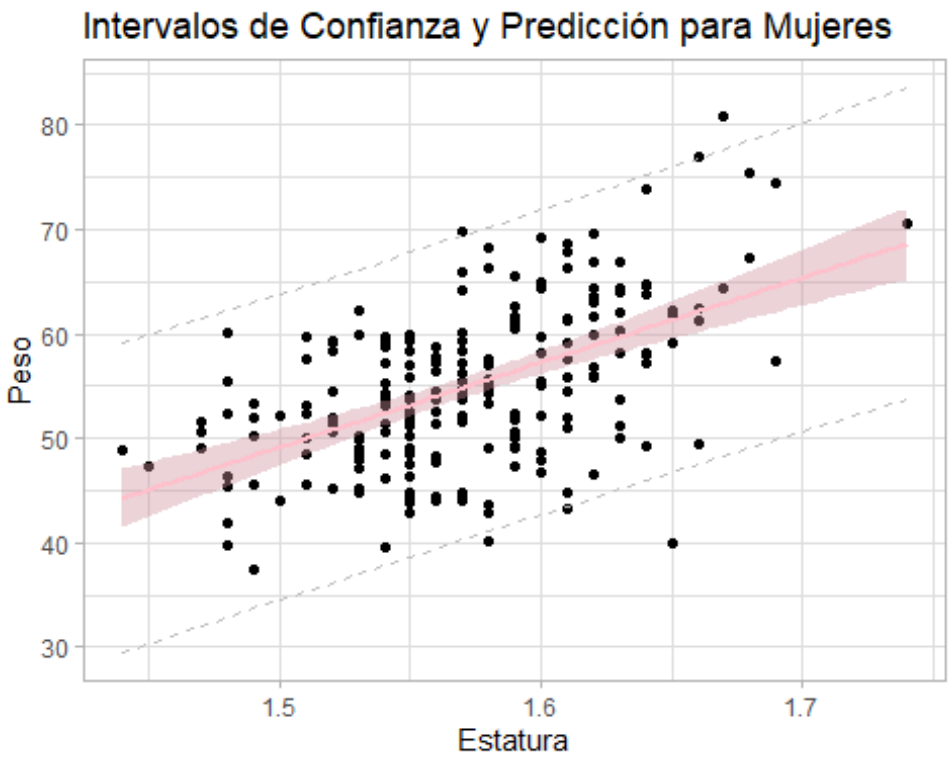
Emite una conclusión final sobre el mejor modelo de regresión lineal que conjunte lo que hiciste en las tres partes de esta actividad.

A pesar de que el modelo 2 no está mal, vamos a agarrar el modelo 1 (Hombres y Mujeres) ya que el modelo de los hombres es el más estable de todos. Aparte que los modelos 1 de hombres y mujeres si tienen Homocedasticidad.

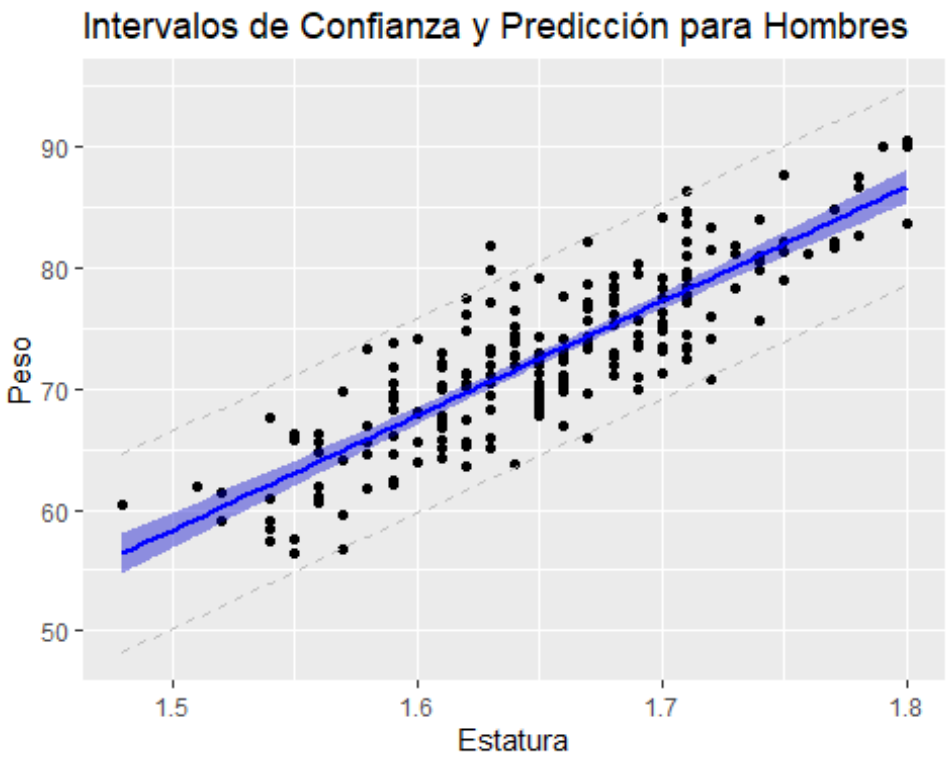
Intervalos de confianza

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.

```
MM <- subset(M, M$Sexo == "M")
MH <- subset(M, M$Sexo == "H")

Modelo1H <- lm(Peso ~ Estatura, data = MH)
Modelo1M <- lm(Peso ~ Estatura, data = MM)

Ip_M <- predict(object = Modelo1M, newdata = MM, interval = "prediction",
level = 0.97)

MM <- cbind(MM, Ip_M)

library(ggplot2)

ggplot(MM, aes(x = Estatura, y = Peso)) +
  geom_point() +  # Puntos observados
  geom_line(aes(y = lwr), color = "gray", linetype = "dashed") +
  geom_line(aes(y = upr), color = "gray", linetype = "dashed") +
  geom_smooth(method = "lm", formula = y ~ x, se = TRUE, level = 0.97,
col = "pink", fill = "pink3") +
  theme_light() +
  labs(title = "Intervalos de Confianza y Predicción para Mujeres",
       x = "Estatura",
       y = "Peso")
```




```r
Ip_H <- predict(object = Modelo1H, newdata = MH, interval = "prediction",
level = 0.97)
MH <- cbind(MH, Ip_H)

ggplot(MH, aes(x = Estatura, y = Peso)) +
  geom_point() +  # Puntos observados
  geom_line(aes(y = lwr), color = "gray", linetype = "dashed") +
  geom_line(aes(y = upr), color = "gray", linetype = "dashed") +
  geom_smooth(method = "lm", formula = y ~ x, se = TRUE, level = 0.97,
col = "blue", fill = "blue3") +
  labs(title = "Intervalos de Confianza y Predicción para Hombres",
       x = "Estatura",
       y = "Peso")
```



Interpreta y comenta los resultados obtenidos

Efectivamente el modelo se ajusta un poco mejor debido a que la parte de los hombres es más estable, el problema es los datos de las mujeres. Podemos solucionar estos problemas con alguna transformación en los datos al momento de hacer los modelos