Actividad 12

Oskar Arturo Gamboa Reyes

2024-09-04

Problema 1

Leer 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)

head(M1)
```

```
##   MH.Estatura MH.Peso MM.Estatura MM.Peso
## 1        1.61   72.21        1.53   50.07
## 2        1.61   65.71        1.60   59.78
## 3        1.70   75.08        1.54   50.66
## 4        1.65   68.55        1.58   56.96
## 5        1.72   70.77        1.61   51.03
## 6        1.63   77.18        1.57   64.27
```

Correlación

```
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 alta correlacion entre el peso y la altura de los hombres con 0.84, aunque menor tambien se observa bastente correlacion entre el peso y la altura de las mujeres con 0.52.

Medidas para analizar

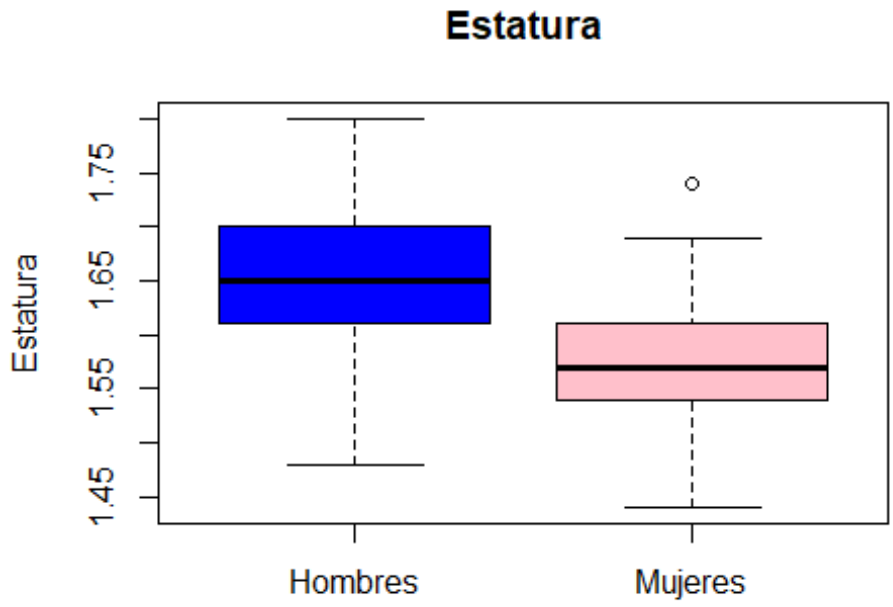
```
n=4 #número de variables
d=matrix(NA,ncol=7,nrow=n)
for(i in 1:n){
  d[i,]<-c(as.numeric(summary(M1[,i])),sd(M1[ ,i]))
}
m=as.data.frame(d)
```

```
row.names(m)=c("H-Estatura","H-Peso","M-Estatura","M-Peso")
names(m)=c("Minimo","Q1","Mediana","Media","Q3","Máximo","Desv Est")
m
```

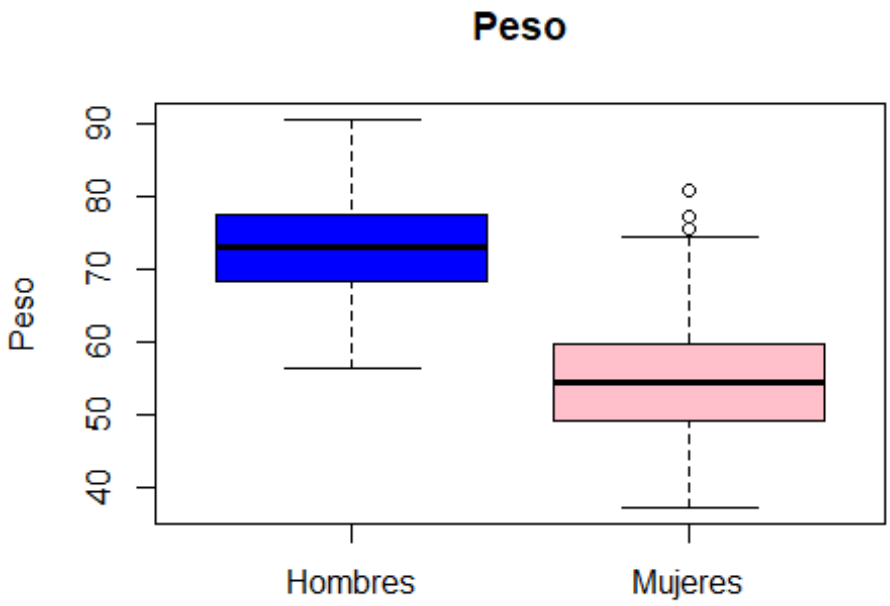
```
##            Minimo      Q1 Mediana     Media      Q3 Máximo   Desv Est
## H-Estatura   1.48  1.6100   1.650  1.653727  1.7000   1.80 0.06173088
## H-Peso      56.43 68.2575  72.975 72.857682 77.5225  90.49 6.90035408
## M-Estatura   1.44  1.5400   1.570  1.572955  1.6100   1.74 0.05036758
## M-Peso      37.39 49.3550  54.485 55.083409 59.7950  80.87 7.79278074
```

Gráficas para datos

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




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



Rectas de mejor ajuste

Dos rectas

```
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:

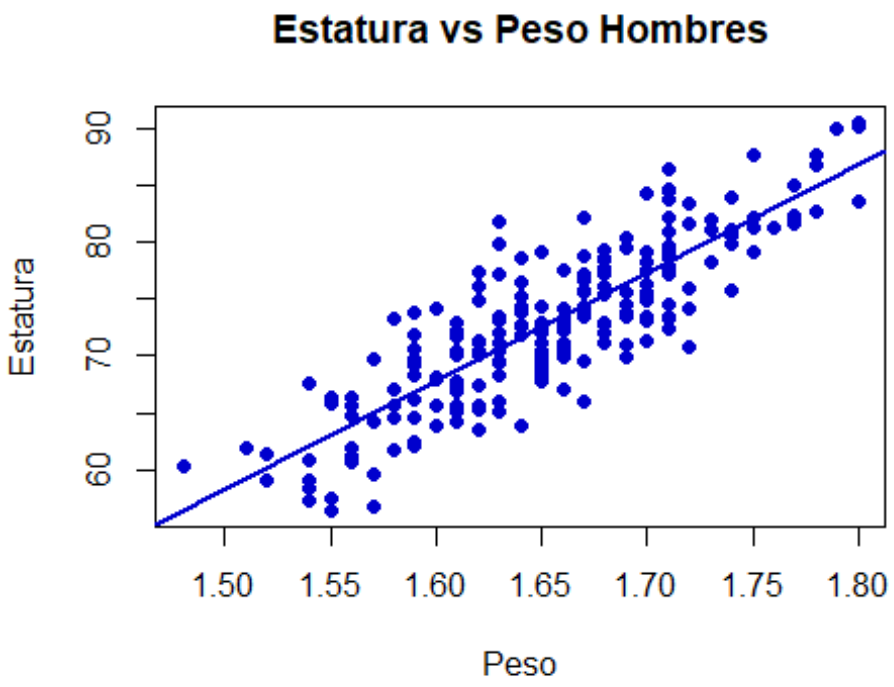
- $H_0: \beta_1 = 0$
- $H_1: \beta_1 \neq 0$

Hombres

```
summary(Modelo1H)
```

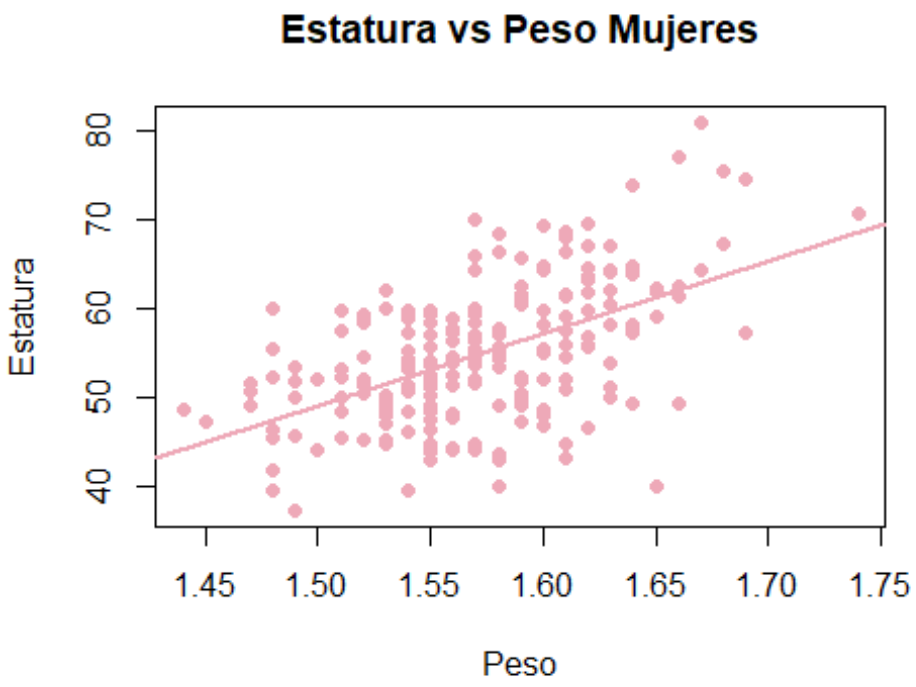
```
##
## Call:
## lm(formula = Peso ~ Estatura, data = MH)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3881 -2.6073 -0.0665  2.4421 11.1883
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -83.685      6.663  -12.56   <2e-16 ***
## Estatura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

```
plot(MH$Estatura,MH$Peso, col="blue3", main="Estatura vs Peso Hombres",
ylab="Estatura", xlab = "Peso",pch=19)
abline(lm(Peso~Estatura, MH), col="blue3",lwd=2)
```



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

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



Un modelo

```
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
```

A 0.05 si es significativo y los modelos quedarían:

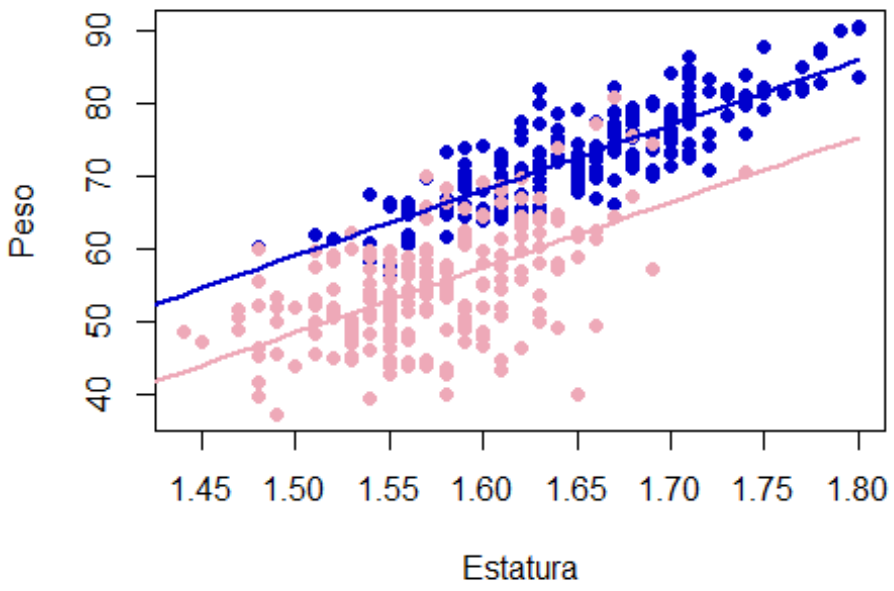
Hombre:

Peso = -74.7546 + 89.2604E

Mujeres:

Peso = -85.3191 + 89.2604E

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

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

colores = c("blue3", "pink2")

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

x= seq(1.40,1.80,0.01)
lines(x, Ym(x), col="pink2",lwd=2)
lines(x, Yh(x), col="blue3",lwd=2)
```



Modelo con interacción

Hipótesis:

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

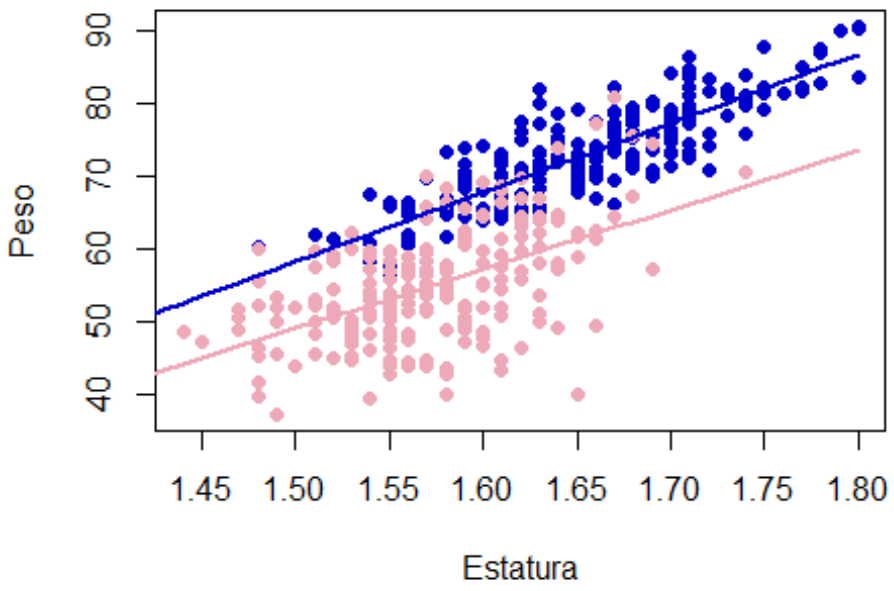
```
Modelo3 = lm(Peso~Estatura*Sexo, M)
Modelo3

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

Peso = −83.6845 + 94.6602Estatura + 11.1241SexoM − 13.5111(Estatura * SexoM)

```
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
```

```r
b0 = Modelo3$coefficients[1]
b1 = Modelo3$coefficients[2]
b2 = Modelo3$coefficients[3]
b3 = Modelo3$coefficients[4]

Yh = function(x){b0+b1*x}
Ym = function(x){b0+b1*x+b2+b3*x}

colores = c("blue3", "pink2")

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

x= seq(1.40,1.80,0.01)
lines(x, Ym(x), col="pink2",lwd=2)
lines(x, Yh(x), col="blue3",lwd=2)
```



Conclusión

¿Qué información proporciona β̂0 sobre la relación entre la estatura y el peso de hombres y mujeres? Interpreta y compara entre este modelo con los 3 modelos anteriores.

Esta beta nos describe la magnitud incial del modelo, en el contexto del problema no tiene mucho sentido ya que una altura de 0 es imposible, sin embargo si describe un valor inicial del peso que afecta a toda la población. Todos los modelos tienen una beta 0 similar, ya que lo que afectan estos modelos es principalmente la pendiente.

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

Estos describen la pendiente de nuestra función, podemos ver que hay una relación diferente entre el peso y estatura de los hombres y de las mujeres, en el modelo anteriore esto no se tomaba en cuenta, lo que hacía a nuestro modelo menos significante. Los primeros dos modelos se ajustan de la misma manera que el último, ya que fueron hechos independientemente.

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.

Podemos notar que el modelo que toma en cuenta la interacción es un mejor modelo que describe mejor el comportamiento de los datos, ya que ahora si existe una variación entre las pendientes de las gráficas que se ajustan dependiendo a sus datos dependiendo de la

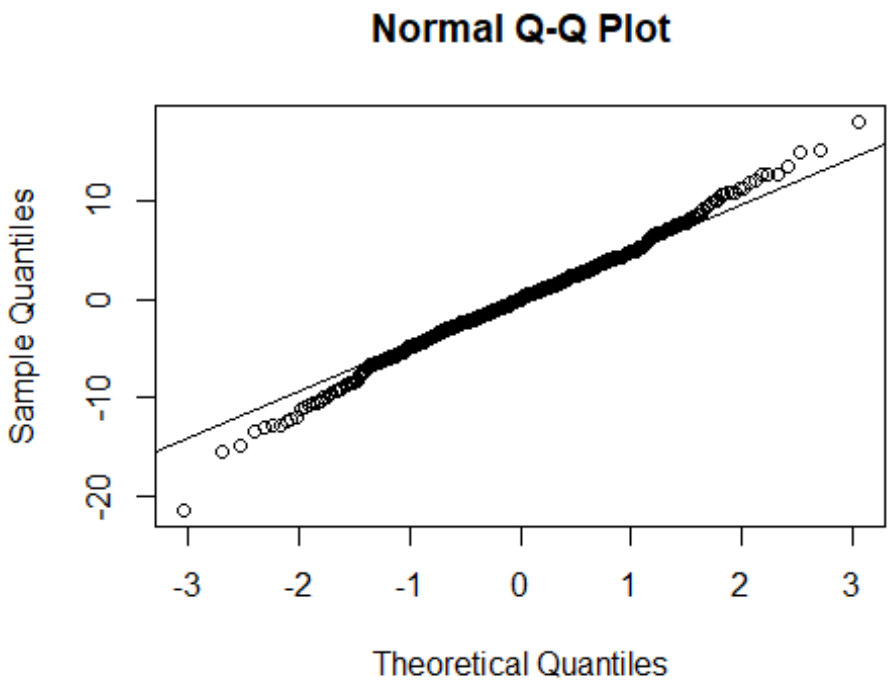
variable sexo. Los dos primeros modelos explican bien el comportamiento pero estan independientes por loq ue no nos dejan apreciar las diferencias entre los sexos. El último modelo es mejor.

Análisis de Errores

Normalidad de los residuos

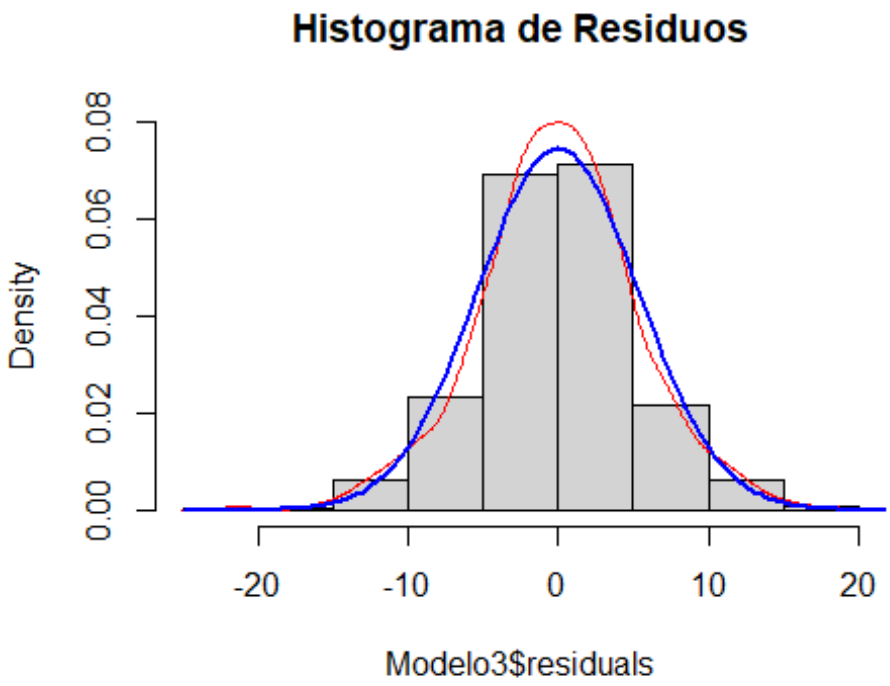
H0: Los datos provienen de una población normal H1: Los datos no provienen de una población normal Regla de decisión: Se rechaza H0 si valor p < 0.03

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

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

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




```
hist(Modelo3$residuals,freq=FALSE, ylim = c(0,0.08), main="Histograma de
Residuos")
lines(density(Modelo3$residual),col="red")
```

```
curve(dnorm(x,mean=mean(Modelo3$residuals),sd=sd(Modelo3$residuals)), from=-
25, to=25, add=TRUE, col="blue",lwd=2)
```



Podemos concluir que la hipótesis se cumple ya que nuestro valor de p es mayor al alpha establecido de 0.03. Además podemos ver que las dos gráficas indican que los residuos se comportan con normalidad, el qqplot nos uestra que los datos se apegan a la linea de normalidad y tambien en el histograma vemos que se apega a la densidad de los datos.

Verificación de media cero

H0: $\mu = 0$ H1: $\mu \neq 0$

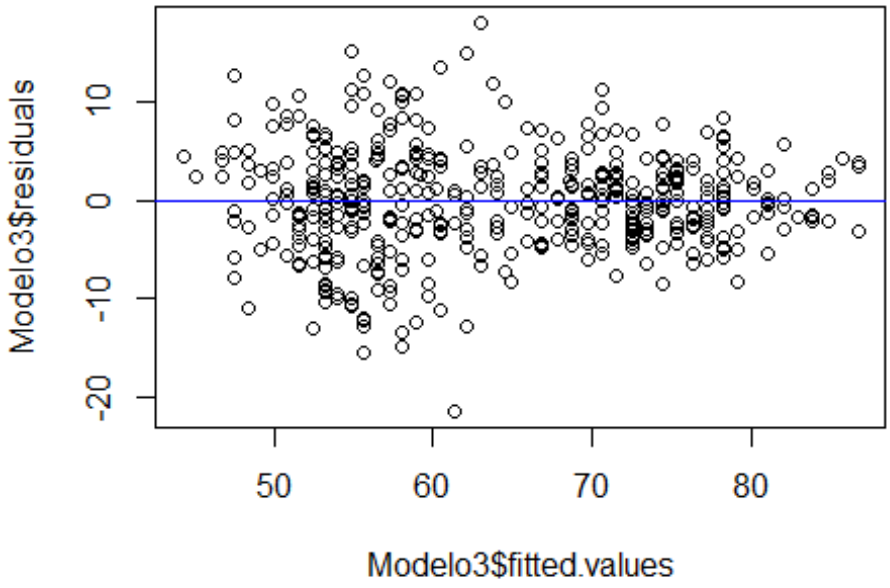
```
t.test(Modelo3$residuals)
```

```
##
##  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
```

Podemos ver que el modelo tiene una media completamente centrada, ya que tenemos un p-value de 1 y nuestros intervalos de confianza son simetricos centrados en 0.

Homocedasticidad

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



Hipótesis

H0: La varianza de los errores es constante H1: La varianza de los errores no es constante

```
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

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

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
```

La hipótesis se va a rechazar ya que podemos notar una pequeña variación en los errores, en los rangos más bajos de peso existe una mayor variación. Esto se debe a que las mujeres tienen una mayor variación en el peso y por eso tenemos una mayor variación en los rangos medios y menores.

Independencia

Hipótesis

H0: Los errores no están correlacionados H1: Los errores están correlacionados

```
dwtest(Modelo3)
```

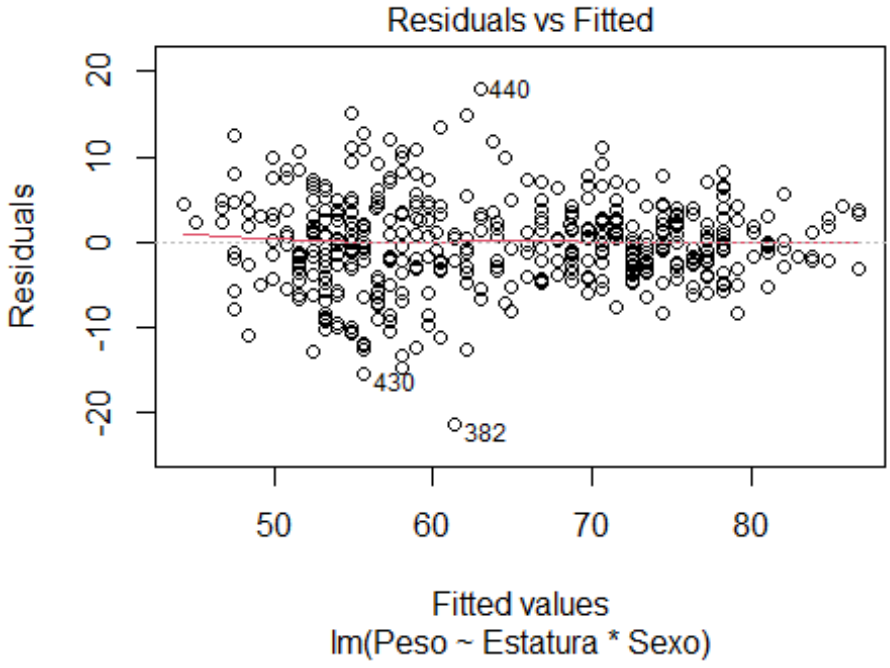
```
##
##  Durbin-Watson test
##
## data:  Modelo3
## DW = 1.8646, p-value = 0.07113
## alternative hypothesis: true autocorrelation is greater than 0
```

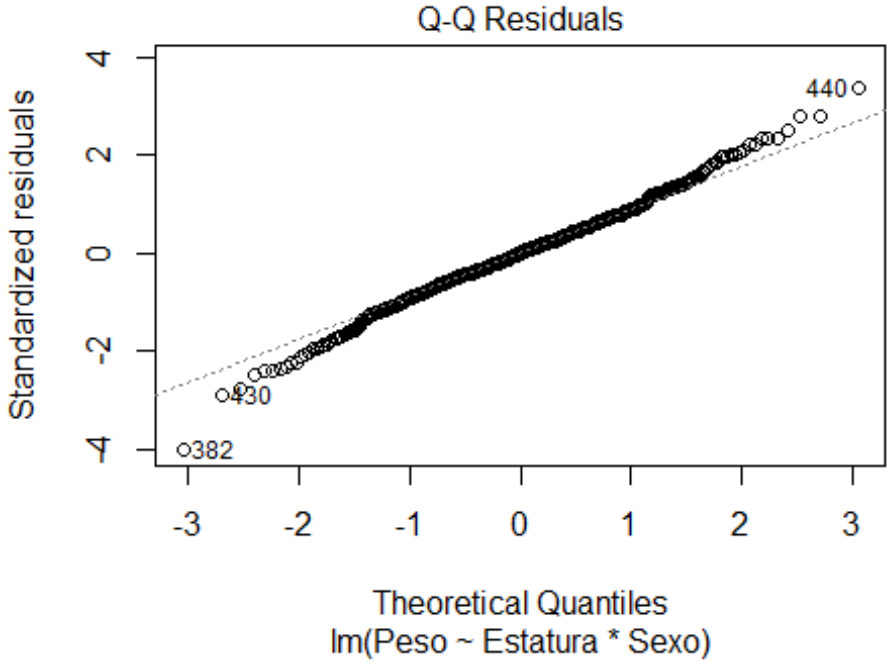
```
bgtest(Modelo3)
```

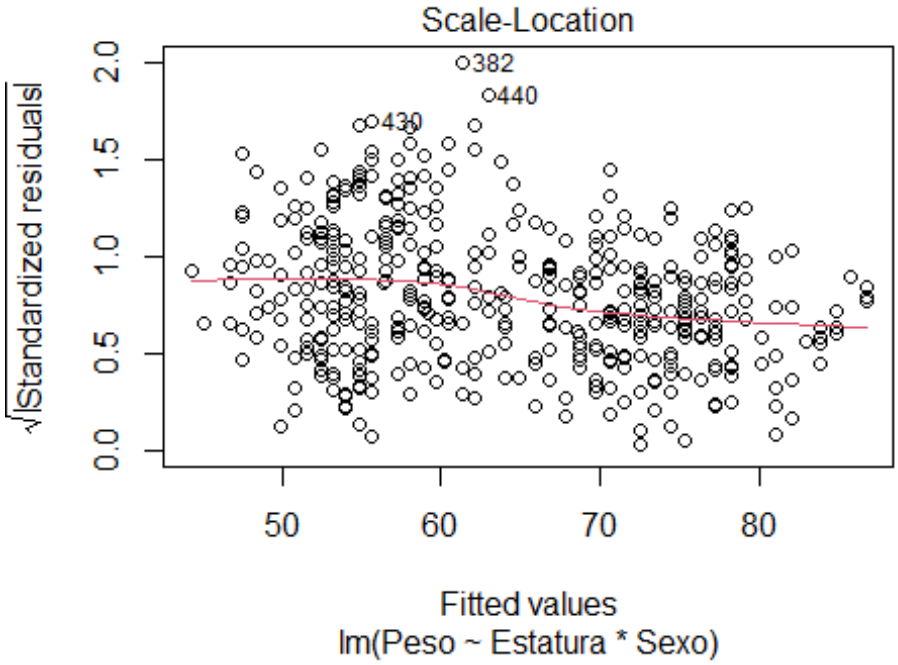
```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  Modelo3
## LM test = 1.3453, df = 1, p-value = 0.2461
```

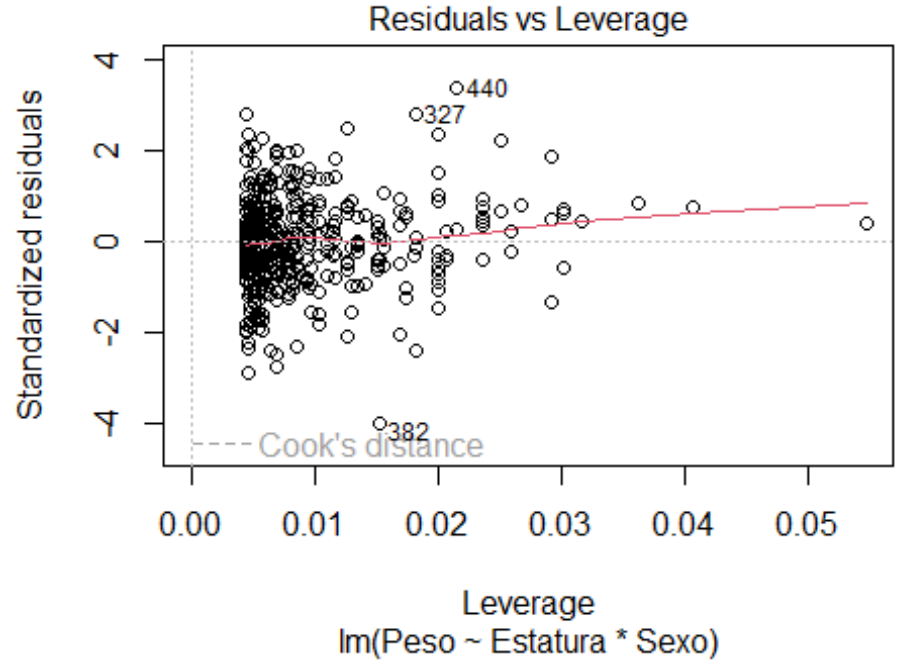
Se puede comprobar la independencia de los residuos ya que tenemos un p-value mayor a 0.03, lo que indica que los errores no están correlacionados.

```
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)





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

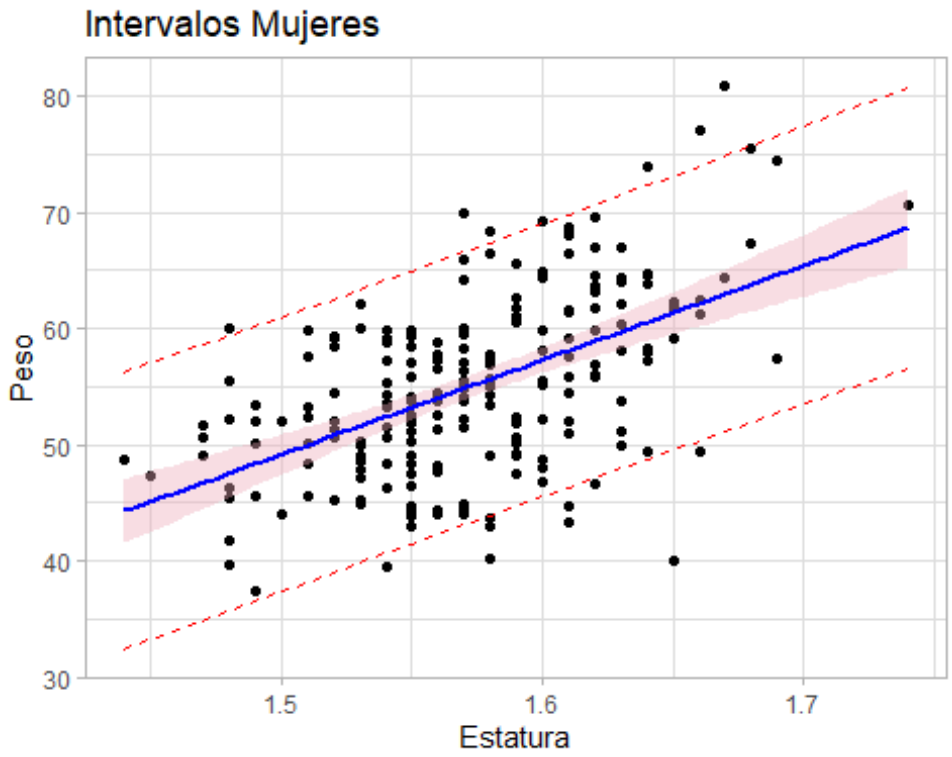
No hay tantas diferencias ya que nos ayudan a analizar la homocedasticidad y la independencia y eso ya lo habiamos analizado con las pruebas y con las graficas anteriores.

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

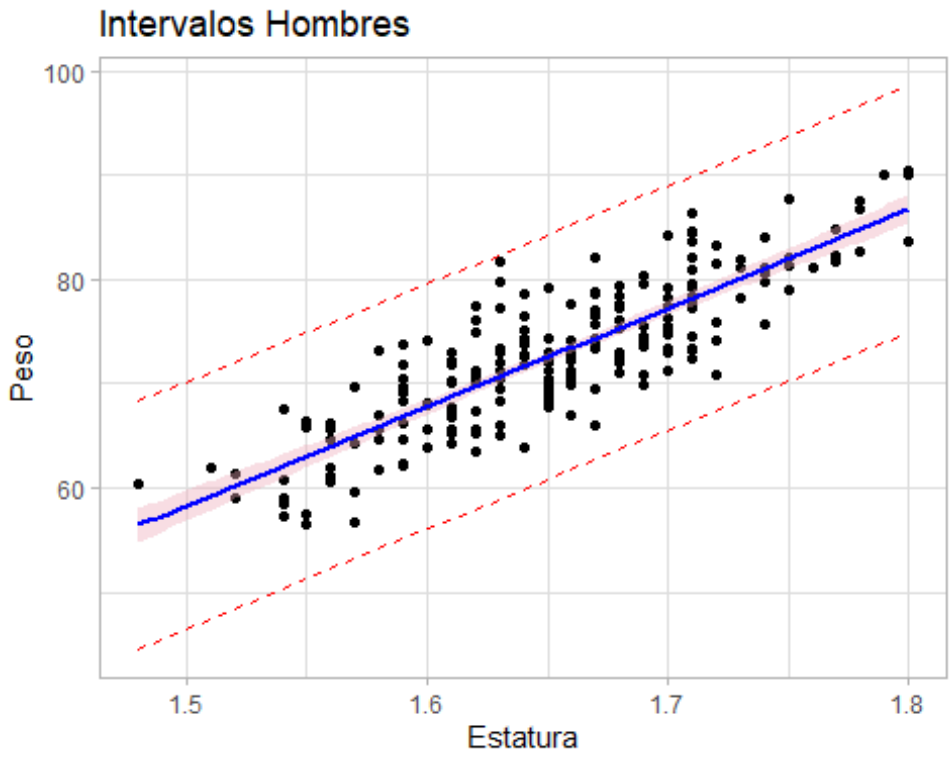
No, sigo pensando que es un buen modelo, cumple con todas las hipótesis lo que nos indica que los errores no son explicados por alguna otra variable.

Intervalos de confianza

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

## Warning in predict.lm(object = Modelo3, interval = "prediction", level =
0.97): predictions 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="blue",
fill="pink2")+
  theme_light()
```



Estas gráficas nos indican los intervalos que podemos esperar de nuestro modelo, como podemos ver las mujeres tienen una mayor variación que no entraría en el intervalo de predicción de nuestro modelo. Por lo que el modelo se ajusta mejor al predecir estaturas de hombres.