12. Regresión Lineal - Análisis de los errores

Eliezer Cavazos

2024-08-30

Analiza la base de datos de estatura y peso de los hombres y mujeres en México y obten el mejor modelo de regresión para esos datos.

```
oPesoEstatura =
read.csv("C:\\Users\\eliez\\OneDrive\\Desktop\\Clases\\Estatura-peso_HyM.csv"
) #Leer la base de datos

oMujeres = subset(oPesoEstatura,oPesoEstatura$Sexo=="M")
oHombres = subset(oPesoEstatura,oPesoEstatura$Sexo=="H")

M1=data.frame(oHombres$Estatura,oHombres$Peso,oMujeres$Estatura,oMujeres$Peso
)
```

La recta de mejor ajuste (Primera entrega)

Analisis Descriptivo

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

```
cor(M1)
```

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

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

```
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

La recta de mejor ajuste

Encuentra la ecuación de regresión de mejor ajuste:

Hombres:

```
ModeloHombres = lm(oHombres$Peso ~ oHombres$Estatura, oHombres)

ModeloHombres
```

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

```
summary(ModeloHombres)
```

```
##
## Call:
## lm(formula = oHombres$Peso ~ oHombres$Estatura, data = oHombres)
##
## 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 ***
## oHombres$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
```

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

Mujeres:

```
ModeloMujeres = lm(oMujeres$Peso ~ oMujeres$Estatura, oMujeres)
ModeloMujeres
```

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

```
summary(ModeloMujeres)
```

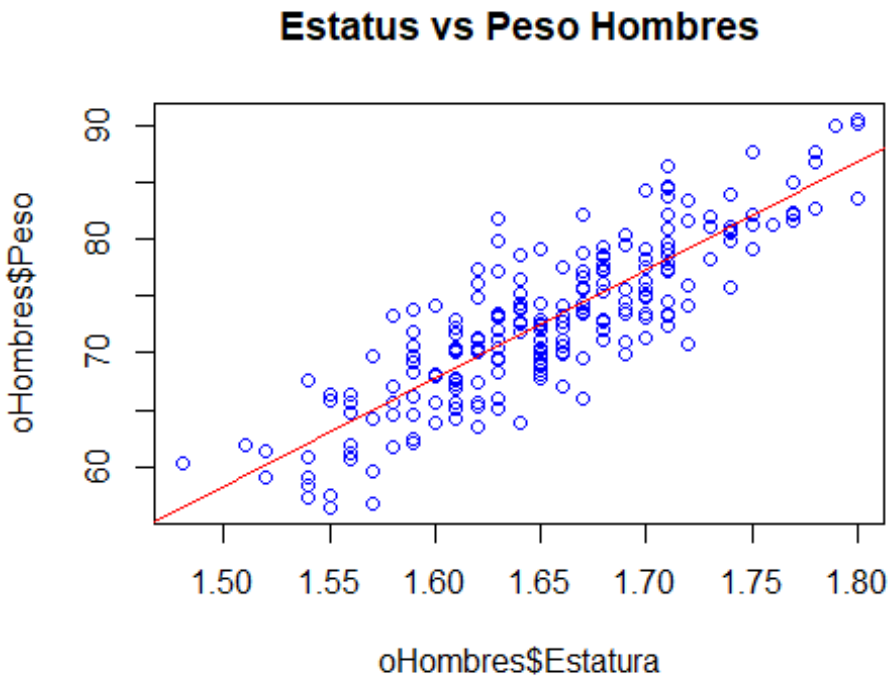
```
##
## Call:
## lm(formula = oMujeres$Peso ~ oMujeres$Estatura, data = oMujeres)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3256  -4.1942   0.4004   4.2724  17.9114
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -72.560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

Realiza la regresión entre las variables involucradas

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

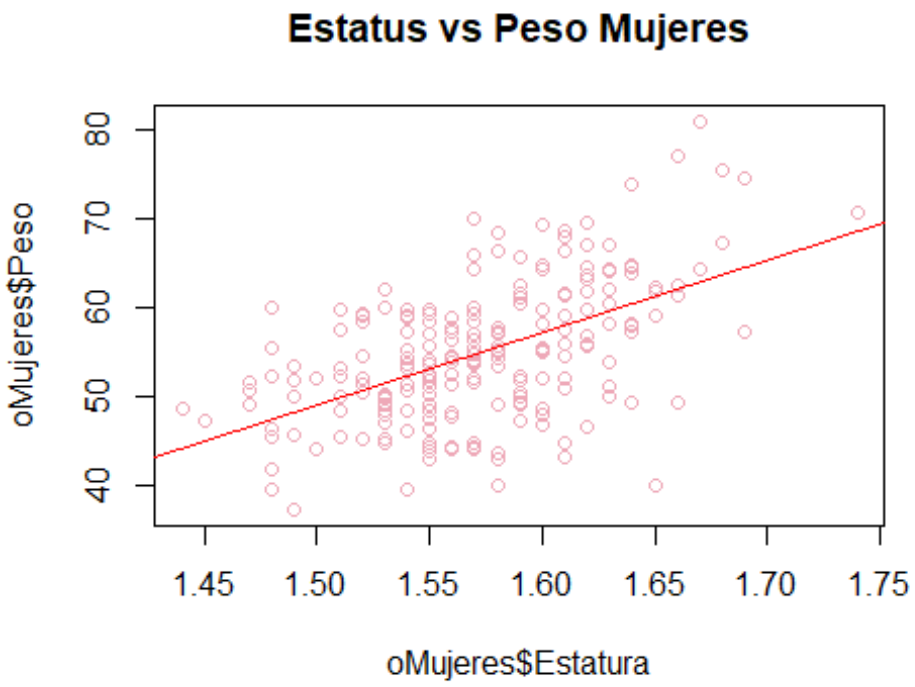
Hombres:

```
plot(oHombres$Estatura, oHombres$Peso, col="blue", main="Estatus vs Peso
Hombres")
abline(ModeloHombres, col="red")
```



Mujeres:

```
plot(oMujeres$Estatura, oMujeres$Peso, col="pink2",main="Estatus vs Peso
Mujeres")
abline(ModeloMujeres, col="red")
```



Verificacion del modelo:

Verifica la significancia del modelo con un alfa de 0.03.

Verifica la significancia de βî con un alfa de 0.03.

Verifica el porcentaje de variación explicada por el modelo

Hombres:

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

summary(Modelo2)

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

```
iNModelo2 = length(Modelo2)
abs( qt(0.03/2,(iNModelo2-2)) )
```

```
## [1] 2.490664
```

A 0.05 si es significativo y los modelos quedarian:

Mujeres: Estatura = -72.56 + 81.15E Estatura = -74.7546 + 89.2604E + -10.5645SexoM Estatura = -85.3191 + 0.0052296E

Hombres: Estatura = -72.56 + 81.15E Estatura = -74.7546 + 89.2604E + -10.5645SexoM Estatura = -74.7546 + 89.2604E

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

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

Ym = function(x){b0+b2+b1*x} #1.2848 + 0.0052296P
Yh = function(x){b0+b1*x} # 1.2727097 + 0.0052296P
colores = c("blue", "pink2")
plot(oPesoEstatura$Estatura,oPesoEstatura$Peso, data=oPesoEstatura,
col=colores[factor(oPesoEstatura$Sexo)],pch=19, ylab="Peso", xlab="Estatura",
main="Relacion Peso vs Estatura")
```

```
## Warning in plot.window(...): "data" is not a graphical parameter
```

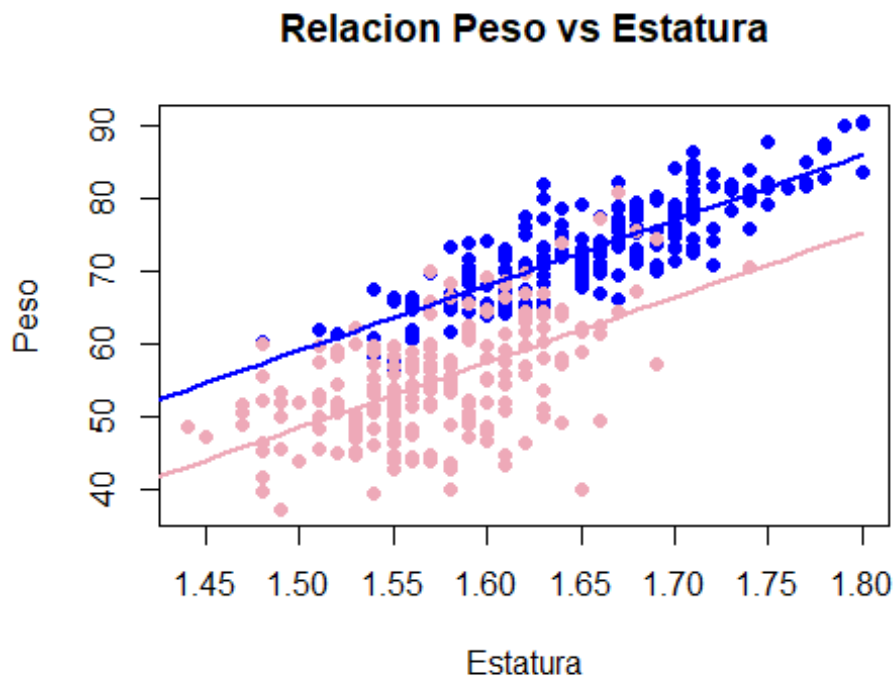
```
## Warning in plot.xy(xy, type, ...): "data" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not
a
## graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not
a
## graphical parameter
```

```
## Warning in box(...): "data" is not a graphical parameter
```

```
## Warning in title(...): "data" is not a graphical parameter
```

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



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

En los anlisis se identifico la relacion del Peso por Estatura tanto en Hombres como Mujeres para poder generar un modelo de regresión lineal y asi poder realizar predicciones para identificar Peso por Estatura ### Interpreta en el contexto del problema: ### ¿Qué información proporciona β0 sobre la relación entre la estatura y el peso de hombres y mujeres? B_0 nos da el valor de Peso cuando la Estatura es 0 y el Sexo es una mujer, aunque esto no tiene mucho sentido cuando la estatura es 0 ya que el modelo esta entrenado sobre estatura mayores a 1.45, con esto dicho no signifique que este mal, solo que al momento de hacer una prediccion dara un mejor resultado mientras la estatura este dentro de un valor "real"/aproximado. ### ¿Cómo interpretas β1 en la relación entre la estatura y el peso de hombres y mujeres? B_1 va a ser el valor que se va a estar sumando(restando en el caso de la formula que tenemos) al valor de B_0 para dar el peso indicado B_1 es tambien multiplicado por la Estatura, ya que siguiendo los datos mientras más Estatura en promedio el peso tambien debe aumentar.

2. Propón un nuevo modelo. Esta vez toma en cuenta la interacción de la Estatura con el Sexo y realiza los mismos pasos que hiciste con los modelos anteriores:

```
oModeloNuevo = lm(Peso~Estatura*Sexo, oPesoEstatura)
summary(oModeloNuevo)

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

```
##
## 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
```

2.1 Obtén el modelo e interpreta las variables Dummy

2.2 Significancia del modelo:

2.2.1 Valida la significancia del modelo con un alfa de 0.03 (incluye las hipótesis que pruebas)

```
iNModeloNuevo = length(oModeloNuevo)

abs( qt(0.03/2,(iNModeloNuevo-2)) )

## [1] 2.490664

oModeloNuevo$coefficients

##    (Intercept)       Estatura          SexoM Estatura:SexoM
##      -83.68454       94.66024       11.12409      -13.51113
```

A 0.03 no es es significativo y los modelos quedarian:

Estatura = -83.68 + 94.66E + 11.124SexoM + -13.511Estatura:SexoM

2.2.2 Validala significancia de βî con un alfa de 0.03 (incluye las hipótesis que pruebas)

β_0 tiene un valor T absoluto de 8.59 que es mayor al valor limite T que es de 2.49 y tambien tiene un valor P menor a 0.03, eso significa que si es significativo.

β_1 tiene un valor T absoluto de 16.092 que es mayor al valor limite de T que es de 2.49 y tambien tiene un valor P menor a 0.03, eso significa que si es significativo

β_2 tiene un valor T absoluto de 0.744 que es menor al valor limite de T y tambien tiene un valor P mayor a 0.03, eso significa que no es significativo

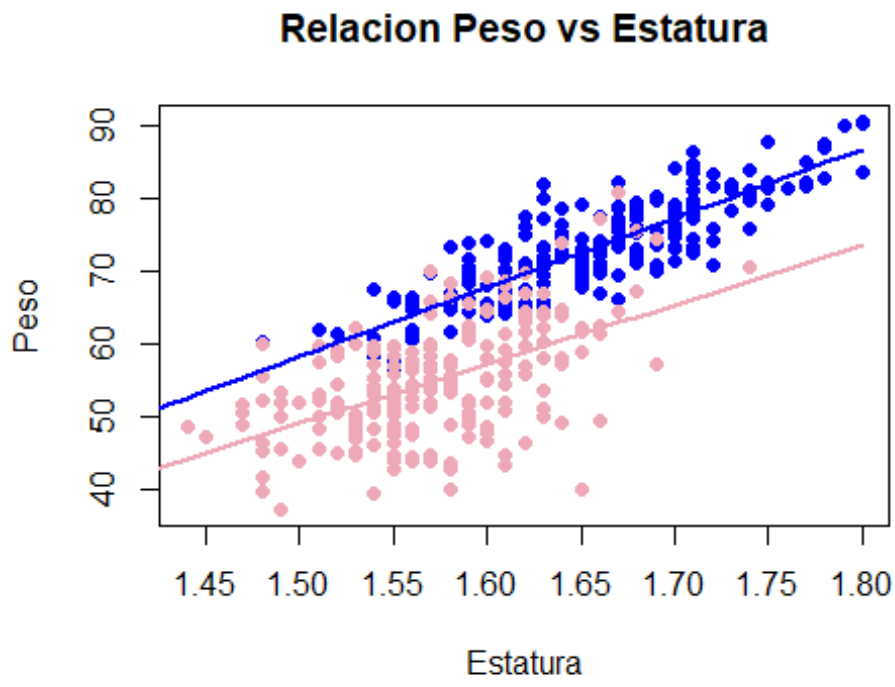
β_3 tiene un valor T absoluto de 1.452 que es menor al valor limite de T y tambien tiene un valor P mayor a 0.03, eso significa que no es significativo

2.2.3 Indica cuál es el porcentaje de variación explicada por el modelo.

El porcentaje de variacion es igual a 78.47%

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

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

Ym = function(x){b0+b2+b1*x+b3*x} #1.2848 + 0.0052296P - 13.51P
Yh = function(x){b0+b1*x} # 1.2727097 + 0.0052296P
colores = c("blue", "pink2")
plot(oPesoEstatura$Estatura,oPesoEstatura$Peso, data=oPesoEstatura,
col=colores[factor(oPesoEstatura$Sexo)],pch=19, ylab="Peso", xlab="Estatura",
main="Relacion Peso vs Estatura")

## Warning in plot.window(...): "data" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "data" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not
a
## graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not
a
## graphical parameter

## Warning in box(...): "data" is not a graphical parameter

## Warning in title(...): "data" is not a graphical parameter

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



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

Este modelo no es bueno porque la interseccion de Sexo por Estatura no es significativo eso significa que en el contexto del problema el mejor modelo es el anterior que realizamos.

3. Interpreta en el contexto del problema:

3.1 ¿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.

B_0 nos da el valor de Peso cuando la Estatura es 0 y el Sexo es una mujer, aunque esto no tiene mucho sentido cuando la estatura es 0 ya que el modelo esta entrenado sobre estatura mayores a 1.45, con esto dicho no signifique que este mal, solo que al momento de hacer una prediccion dara un mejor resultado mientras la estatura este dentro de un valor "real"/aproximado.

Este valor entre los diferentes modelos tiene un cambio dependiendo del Sexo por eso en este ultimo modelo cuando hacemos la combinacion de ambos tiene un valor absoluto más grande

3.2 ¿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.

B_1 va a ser el valor que se va a estar sumando(restando en el caso de la formula que tenemos) al valor de B_0 para dar el peso indicado B_1 es tambien multiplicado por la

Estatura, ya que siguiendo los datos mientras más Estatura en promedio el peso tambien debe aumentar.

Este valor tiene el mismo uso en todos los modelos

B_2 va a ser el valor que se va a estar restando al valor de B_0 cuando sea Mujer por como fue hecho el modelo en este caso cuando se calcule el peso de una mujer va a tener un valor más bajo al valor de peso que puede dar de un Hombre.

B_3 es la combinacion de Estatura y Sexo, este valor no se encuentra en ninguno de los modelos más que en el ultimo donde se usa esta combinacion para sacar el peso combinando Estatura y Sexo pero al final pudimos identificar que no es significactivo

3.3 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 más apropiado es el segundo modelo ya que junta los valores más significativos para dar una predicción del peso más precisa

La Validez del modelo

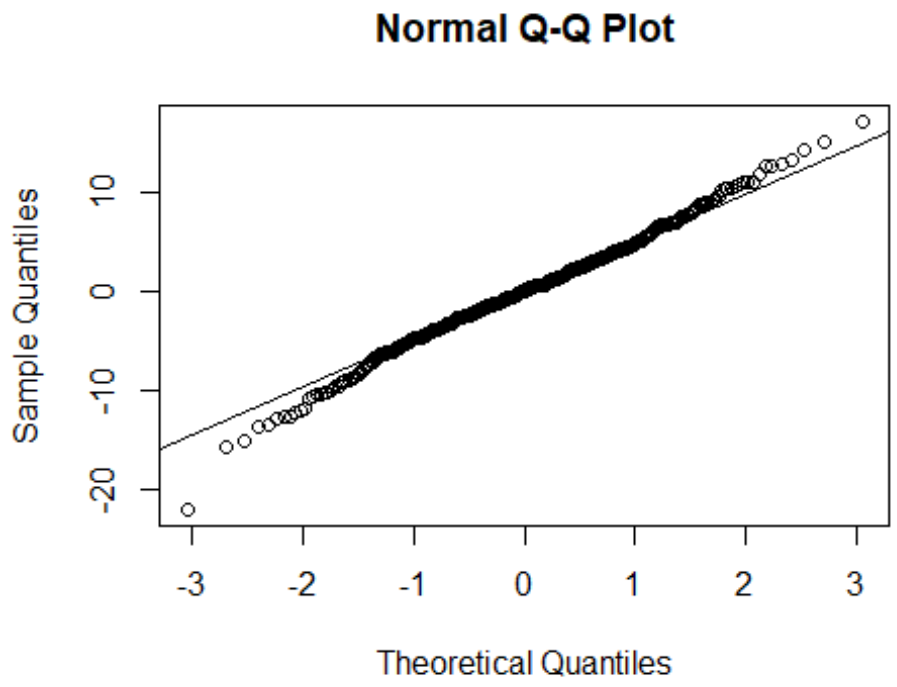
1. Analiza si el (los) modelo(s) obtenidos anteriormente son apropiados para el conjunto de datos. Realiza el análisis de los residuos:

1.1 Normalidad de los residuos

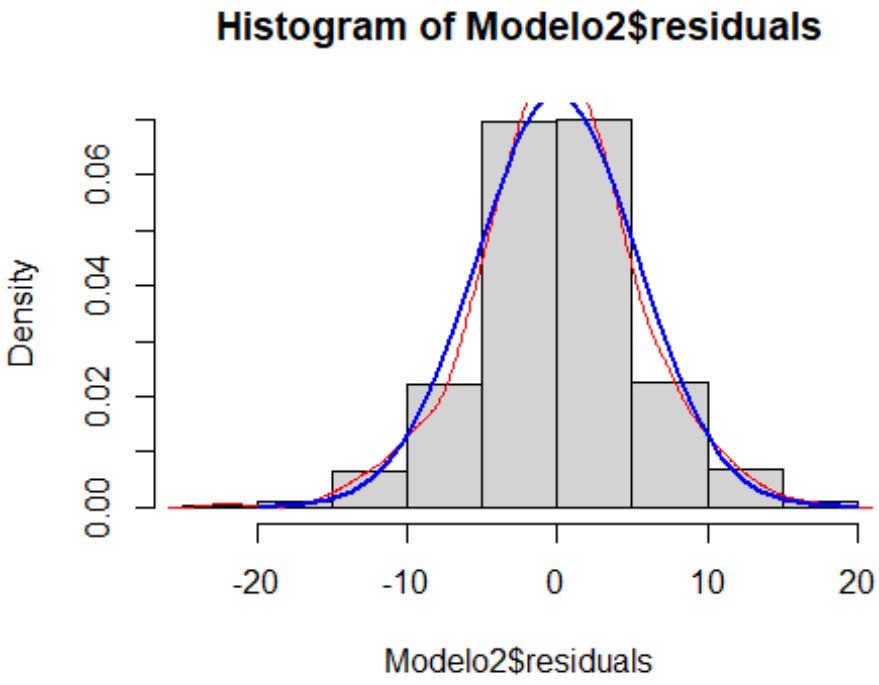
```
library(nortest)
ad.test(Modelo2$residuals)
```

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

```
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=-
20, to=20, add=TRUE, col="blue",lwd=2)
```



1.2 Verificación de media cero

```
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
```

1.3 Homocedasticidad e independencia

Homocedasticidad:

```
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
```

```
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
```

Independencia:

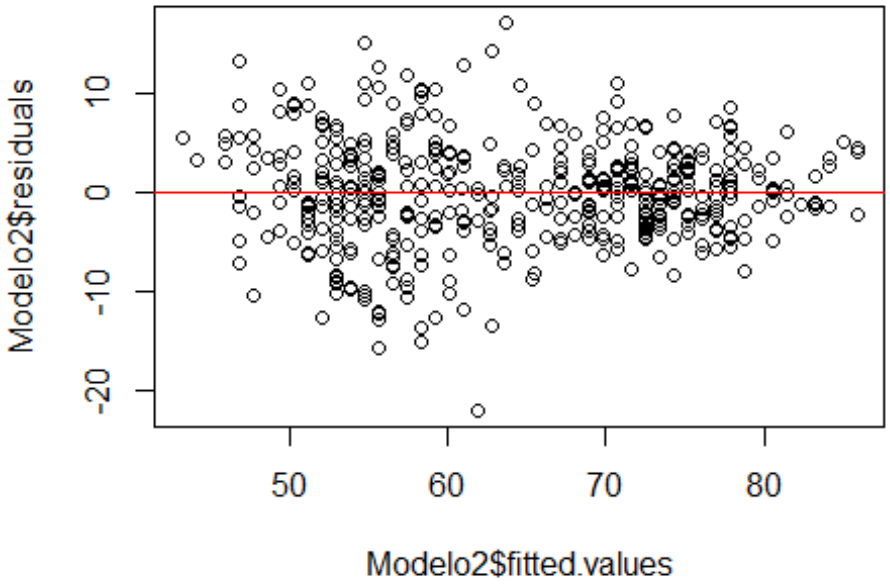
```
dwtest(Modelo2)
```

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

```
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
```

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



El Grafico muestra homocedasticidad y simetria

2. No te olvides de incluir las hipótesis en la pruebas de hipótesis que realices.

En el caso donde nuestra α tiene un valor de 0.05 las hipotesis quedarian

Homocedasticidad: El valor de P es menor a Alpha asi que la Hipotesis incial se rechaza y ademas como la varianza de errores no es constante se mantiene la Hipotesis alternativa/H1.

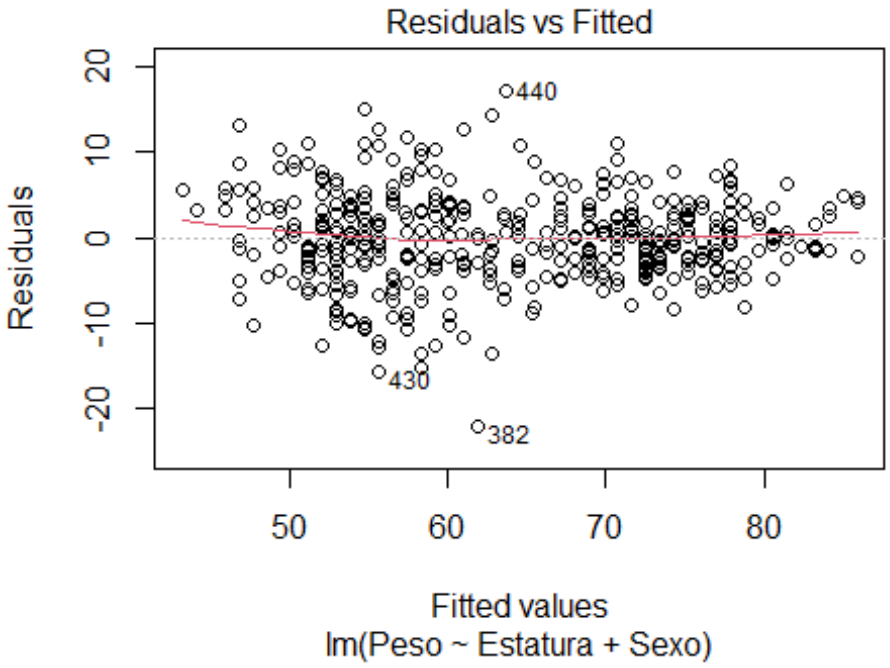
Independencia: El valor de P es menor a Alpha asi que la Hipotesis inicial no se rechaza, así que se identifica que los valores no estan autocorrelacionados

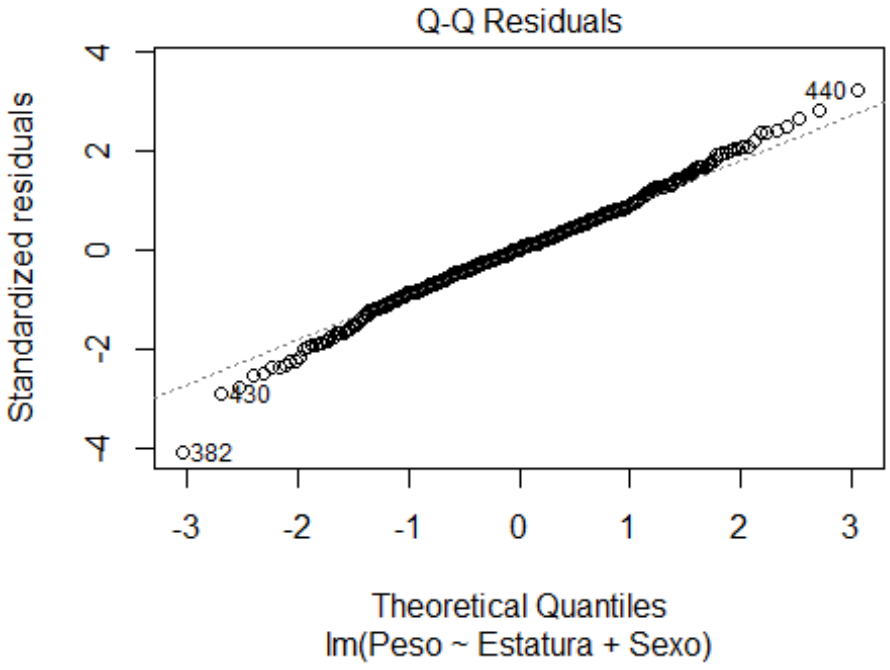
3. Interpreta en el contexto del problema cada uno de los análisis que hiciste.

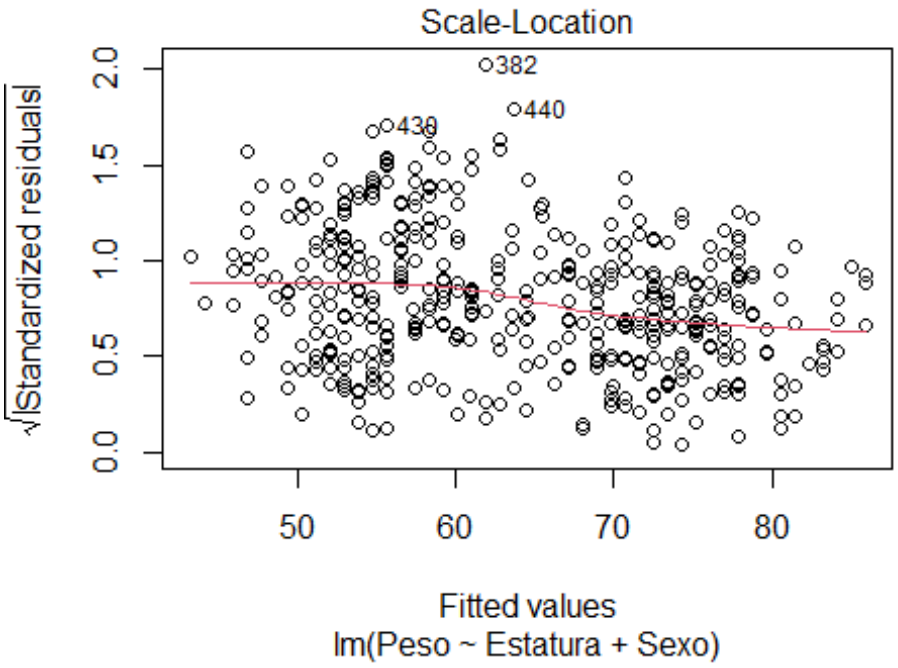
En los analissi que realice se demuestra que los datos tienen independencia, asi que podemos identificar que la independencia podria estar por el valor de sexo, donde dependiendo del sexo los Pesos de Hombres y Mujeres no se autocorrelacionan.

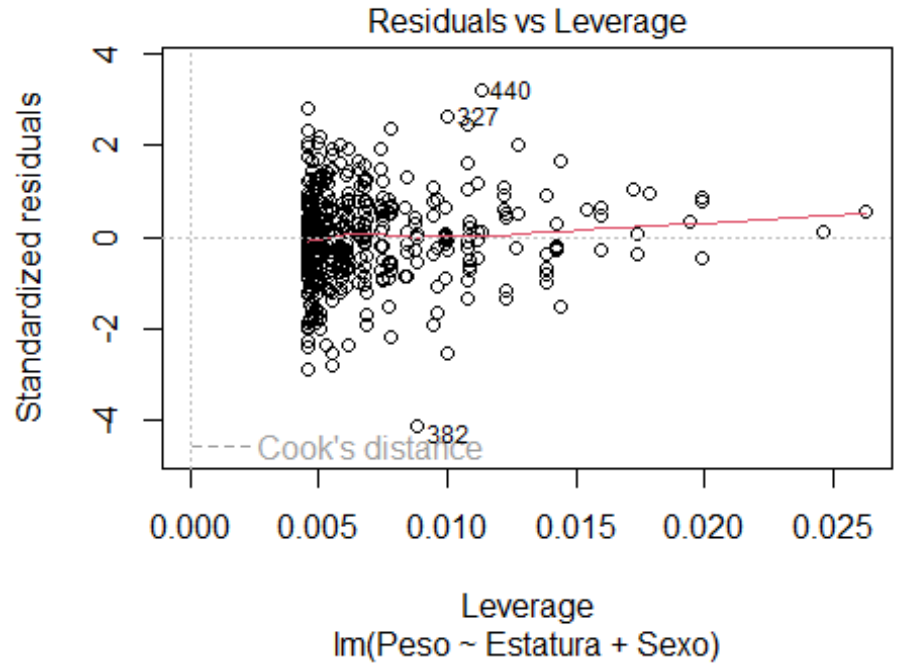
4. Utiliza el comando: plot(modelo). Observa las gráficas obtenidas y contesta:

```
plot(Modelo2)
```


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|
382
440
430
0.0
0.5
1.0
1.5
2.0
50
60
70
80
Fitted values
lm(Peso ~ Estatura + Sexo)


Residuals vs Leverage
Standardized residuals
440
327
382
Cook's distance
4
2
0
-2
-4
0.000
0.005
0.010
0.015
0.020
0.025
Leverage
lm(Peso ~ Estatura + Sexo)

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

Las grafica de QQPlot quedan iguales, mientras que las graficas de residuos tienen la linea roja un poco más distorisionada en esta ultima grafica

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

No cambian las conclusiones que habia obtenido.

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

Al final como ya habiamos identificado el mejor modelo para regresión lineal es el Modelo 2, ya que podemos ver por las graficas de residuos muestra homocedasticidad y simetria, ademas que la grafica de QQPlot muestra que los datos estan Normalizados.

Intervalos de confianza

1. 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.

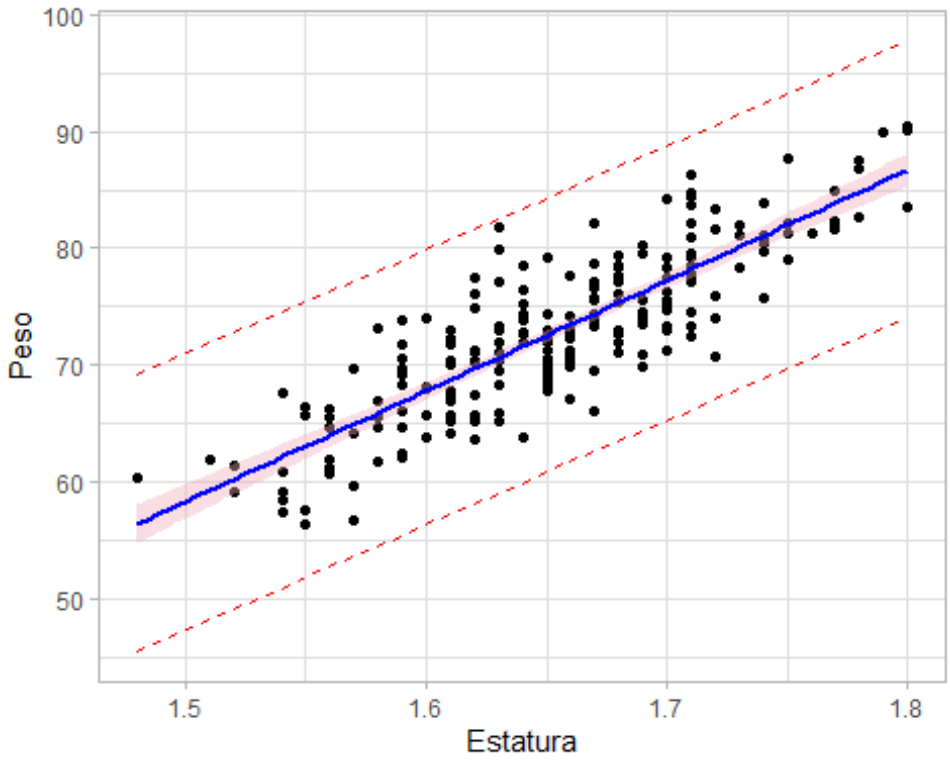
```
Ip=predict(object=Modelo2,interval="prediction",level=0.97)
```

```
## Warning in predict.lm(object = Modelo2, interval = "prediction", level =
0.97): predictions on current data refer to _future_ responses
```

```
M2=cbind(oPesoEstatura,Ip)
M2H = subset(M2, Sexo=="H")
M2M = subset(M2, Sexo=="M")
```

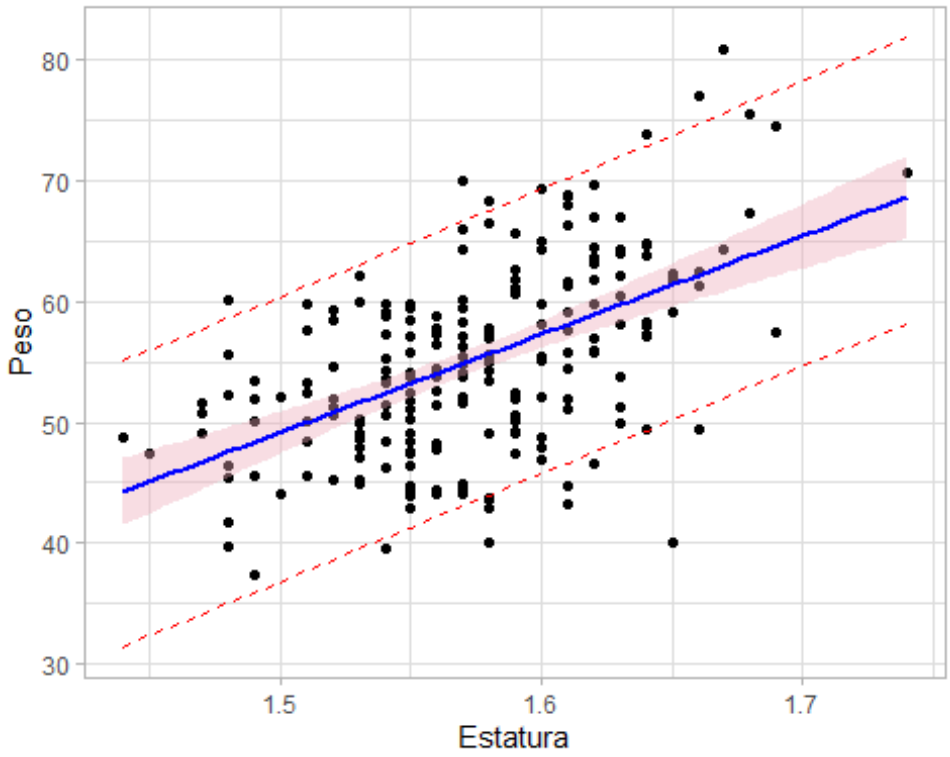
Hombres:

```
library(ggplot2)
ggplot(M2H,aes(x=Estatura,y=Peso))+
geom_point()+
geom_line(aes(y=lwr), color="red", linetype="dashed")+
geom_line(aes(y=upr), color="red", linetype="dashed")+
geom_smooth(method=lm, formula=y~x, se=TRUE, level=0.97, col="blue",
fill="pink2")+
theme_light()
```



Mujeres:

```
library(ggplot2)
ggplot(M2M,aes(x=Estatura,y=Peso))+
geom_point()+
geom_line(aes(y=lwr), color="red", linetype="dashed")+
geom_line(aes(y=upr), color="red", linetype="dashed")+
geom_smooth(method=lm, formula=y~x, se=TRUE, level=0.97, col="blue",
fill="pink2")+
theme_light()
```



2. Interpreta y comenta los resultados obtenidos

Por las graficas podemos identificar las predicciones que se realizan sobre Hombres y Mujeres, donde en la grafica de hobres los intervalos de prediccion quedan dentro de todos los datos mientras que en las mujeres quedan algunos datos atipicos fuera de los intervalos de prediccion.